SNPAnalyzer Pro

Version 1.1

User's Guide

ISTECH
Beyond Bioinformatics

Before using the product

Thank you for choosing our product.

This user's guide provides information about user's direction, installation guide, and operation guide.
Please read this guide before using the product in order to use it appropriately.

- This user's guide and product are protected by the copyright law.
- It is prohibited to copy, reproduce, or translate the part of or whole guide without a prior written permission of ISTECH.
- This product and user's guide may contain printing or technical errors and are subject to change without prior notice.
- ISTECH is not responsible for any damage caused by the use of product for purposes other than those for which it was intended in this guide.

Please read this guide thoroughly before using the product.

<Customer Support>
We listen to our customers. If you have any questions, please contact us in the following ways:
Telephone: 031) 903 – 1155
Fax: 031) 903 – 1152
Email for inquiries: snpanalyzer@istech21.com
Website: http://istech21.com/snpanalyzer
Technical Support: http://istech21.com/

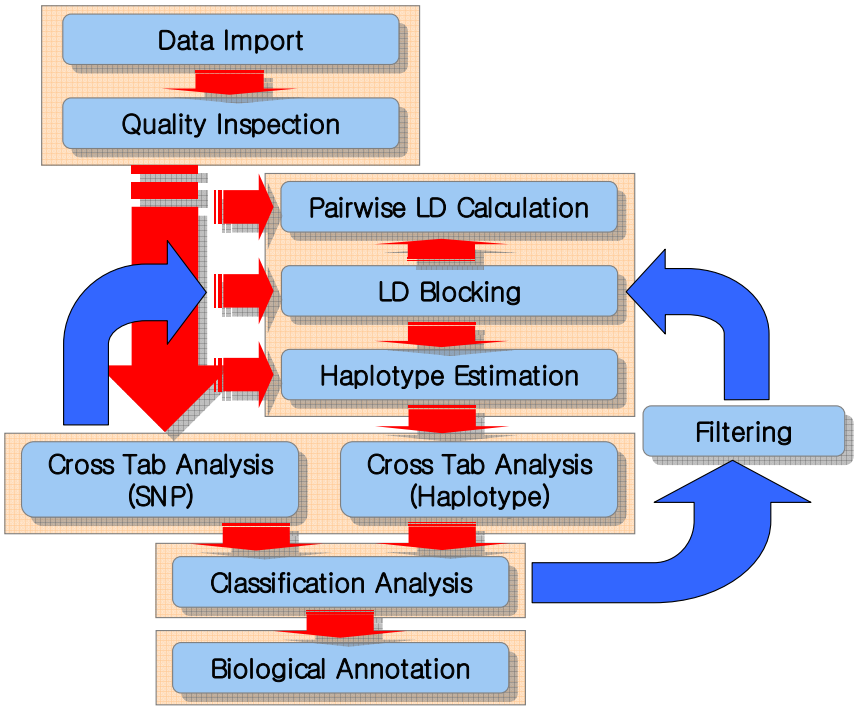
Chapter 1

Introduction

1. Introduction of SNPANALYZER PRO

1.1. Summary

SNP (Single Nucleotide Polymorphism) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species. SNP forms over 90% of the variations occurring in the human genome. In general, the variation occurs from one in 200 to 1000 nucleotides. It is known that the pattern of variation differs from geographical or ethnic groups as well as individuals. Therefore, by analyzing the pattern of SNP occurrence, it provides the foundation to analyze the cause of the difference in sensibility to diseases and reaction to drugs in the DNA sequence level. Although genotyping was performed on a small number of SNPs for a specific gene in the past, the genome-wide SNP chip technology that can simultaneously perform genotyping on from 10,000 to 1 million SNPs distributed in the entire genome is used in a variety of biological research. SNPAnalyzer-Pro is a SNP analysis specialty program that can analyze not only a small scale of SNP genotype data but also a large scale of genome-wide SNP chip data in various ways through the user-friendly interface. SNPAnalyzer-Pro can perform various analyses: case-control study, classification analysis, and Linkage Disequilibrium relationship analysis and it provides a variety of biological annotation information in real-time.



<Figure 1-1> SNP Analysis Process using SNPAnalyzer-Pro

1.2. Main Functions

1.2.1. Data Import

Genotype data is entered and converted into the type that can be used in a variety of analysis later. Main genotype types that can be analyzed using SNPAnalyzer-Pro are the following:

- Affymetrix GeneChip® Genotype Series
- Illumina Infinium Whole-Genome Genotyping Assay
- ABI TaqMan® Genotyping Assays

※ Maximum number of SNPs for analysis: Over 500,000 (500K)
※ Maximum number of samples for analysis: Over 2000
※ Genotype data of Affymetrix are test files created by GCOS/GTYPE or DTT/Genotyping Console software.
※ Genotype data of Illumina and ABI are text files created by BeadStudio and SDS software respectively.

1.2.2. PreProcess

It filters unusable SNPs and samples or substituting empty data into appropriate values for the entered genotypes through a variety of methods. The preprocessing methods are the following:

- Remove Sample by Genotype Call Rate
- Remove Monomorphic SNP
- Remove SNP by Minor Allele Frequency
- Remove SNP by HWE (Hardy-Weinberg Equilibrium) Test
- Replace Missing Genotype with Appropriate Value

1.2.3. Cross Tabulation Analysis

Cross Tabulation Analysis extracts SNPs and haplotypes that show statistically significant difference in allele frequency or genotype frequency observed in the case and control samples. Particularly, for the analysis using SNPs, it performs a maximum of 10 analyses considering risk factor and genetic model. Additional analysis results are OR (Odds Ratio), AR% (Attributable Risk %), and PAR% (Population Attributable Risk %).

- Risk Factor

- Minor Allele / Major Allele

- Genetic Model
  - Additive Model
  - Codominant Model
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Estimated Value
  - Odds Ratio
  - Attributable Risk %
  - Population Attributable Risk %

1.2.4. Logistic Regression Analysis

It extracts the most suitable SNPs and haplotypes to discriminate case and control samples by applying dichotomy logistic analysis model and forward variable selection. Like cross tabulation analysis, it performs a maximum of 10 analyses.

- Risk Factor
  - Minor Allele / Major Allele
- Genetic Model
  - Additive Model
  - Codominant Model
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Estimated Value
  - Parameter Estimation for Classification Feature
  - Classification Table

1.2.5. LD (Linkage Disequilibrium) Analysis

SNPs located relatively close in a genome show strong Linkage Disequilibrium and haplotype sequence can be estimated using theses SNPs in strong Linkage Disequilibrium. Also, it calculates the tagging SNPs that represent a number of SNPs.

- Pairwise LD Calculation
- Tagging SNPs Selection
- LD Blocking

- Haplotype Estimation
- Crossover Rate Calculation

1.2.6. Biological Annotation

Biological annotation information is automatically extracted from significant SNPs extracted from cross tabulation analysis, logistic regression analysis, and LD analysis. Especially, it provides the information of genes in which SNPs are located, and gene ontology along with a variety of SNP annotation information provided from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) of NCBI.

- SNP Annotation
  - Physical Position
  - Functional Information: Non-Synonymous, Synonymous, Intron, Locus Region, Up/Down Stream
- Gene Annotation
  - Gene Symbol, Gene ID, GO ID, GO Term, Category

1.3. Recommended Specifications

- Minimum Specifications
  - OS: Microsoft Windows 2000/XP System (internet connection required)
  - CPU: Pentium 4 2.4GHz or higher
  - RAM: 1GB or more
  - Storage: Over 2GB on installation (separate genotype data storage space required)
- Required Application Program
  - J2SE Runtime Environment 5.0 or higher (installed with SNPAnalyzer-Pro)

Chapter 2

Data Import

2. Data Import

It is the process of converting genotype into data format that can be later used in a variety of analysis. Once completes data input, it automatically performs data PreProcess according to the option configured in default. After completing PreProcss, Annotation files for entered SNPs are automatically created. Genotype data, PreProcess result data, and annotation data are all added in a newly created project for the user to view right away.

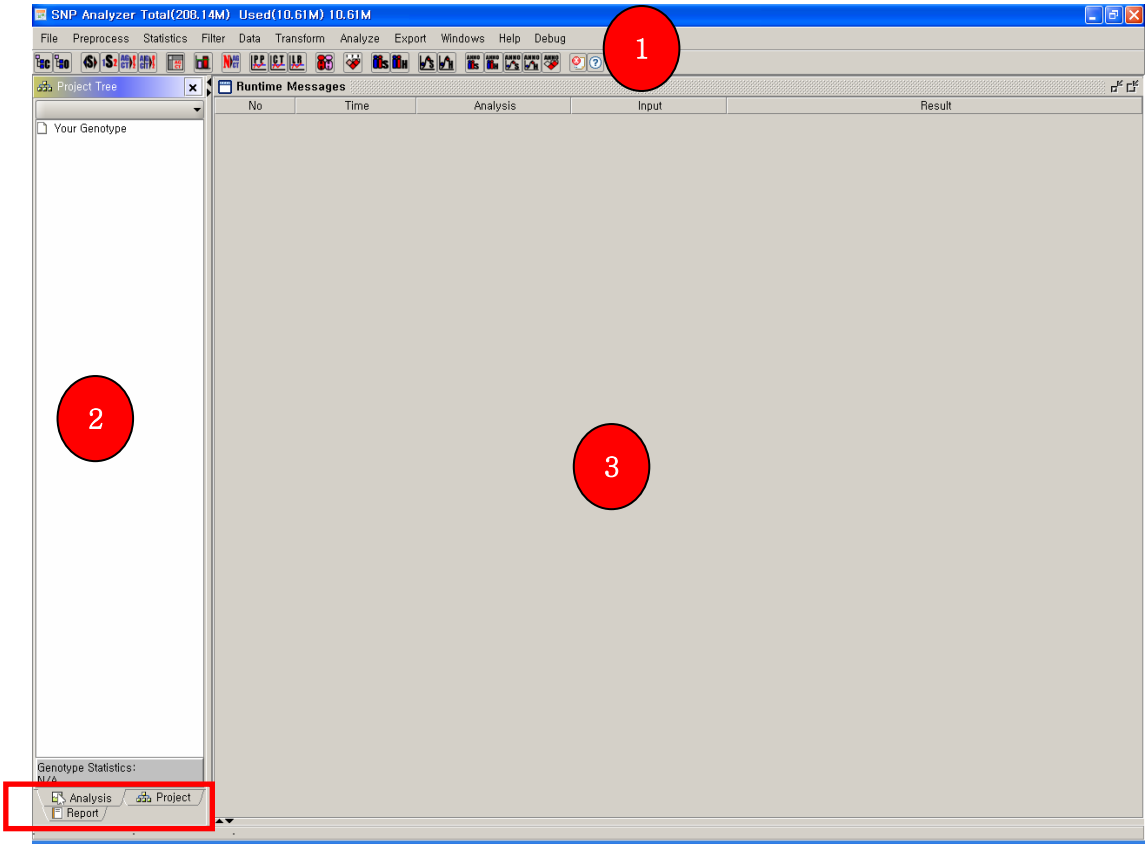
2.1. File

When you run SNPAnalyzer-Pro, a screen like <Figure 2-1> appears.

① Analysis-related Menu and Hot Key area

② Input data and project configuration data tree structure area (3 tabs: Analysis, Project, and Report)

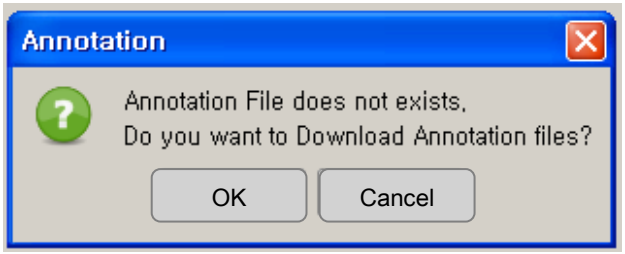
③ Analysis process status display area



<Figure 2-1> SNPAnalyzer-Pro initial screen

When you run the program without SNP annotation file and gene annotation file installed with SNPAnalyzer-Pro, a pop-up window shows to automatically download the files as in

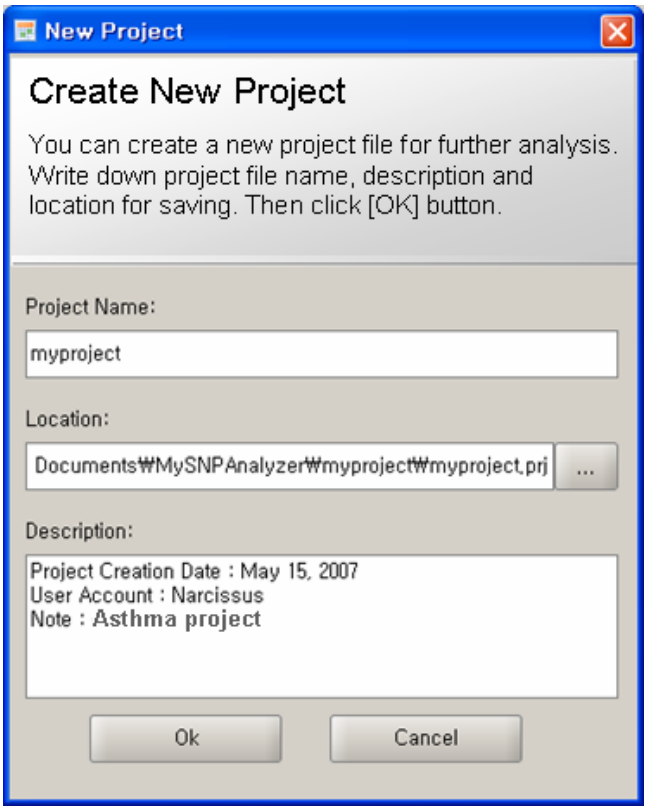
<Figure 2-2>. Click [OK] to download annotation file. It takes 1 to 5 minutes to complete the download depending on your Internet connection.



<Figure 2-2> Annotation file download notification window

2.1.1. Create and Import Project File

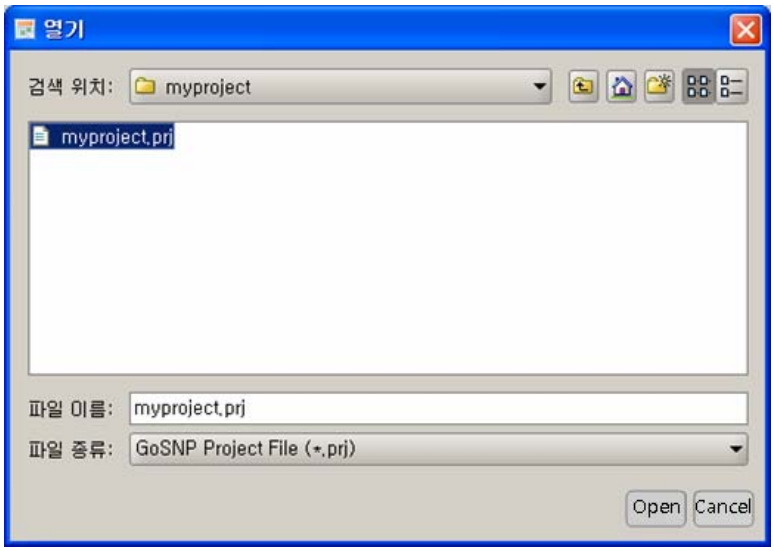
First you need to create a new project or import an existing project in order to analyze SNP data. Click [File] and then the [New Project] menu to show the screen in which you can create a new project as in <Figure 2-3>. After entering a project name in "Project Name" and simple description of the project in "Description", and click [OK] to create a new Project.



<Figure 2-3> Create new project

Click [File] > [Open Project] to display the screen as in <Figure 2-4>. Click [Open] after

selecting a project file in order to import an existing project.



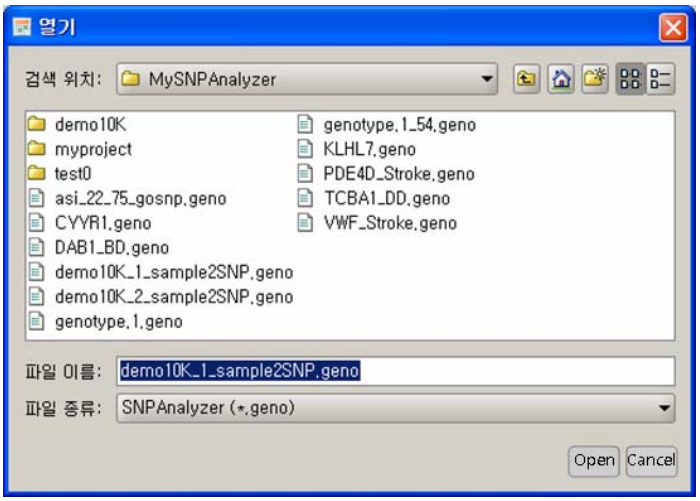
<Figure 2-4> Import existing project

2.1.2. Save and Close Project

Click [File] > [Save Project] to save project and click [File] > [Close Project] to close project.

2.1.3. Import Genotype Data (SNPAnalyzer-Pro Format)

In the main menu, click [File] > [Import Data] > [SNPAnalyzer Format (SNP To Sample)] > [With SNP Annotation] or [File] > [Import] > [SNPAnalyzer Format(SNP To Sample)] > [Without SNP Annotation] or [File] > [Import Data] > [SNPAnalyzer Format (Sample To SNP)] > [With SNP Annotation] or [File] -> [Import Data] > [SNPAnalyzer Format (Sample To SNP)] > [Without SNP Annotation] to show the window where you can select genotype data. Click [Open] after selecting genotype file. For more information on input data format and related topics, please refer to Chapter 7, Data Format.



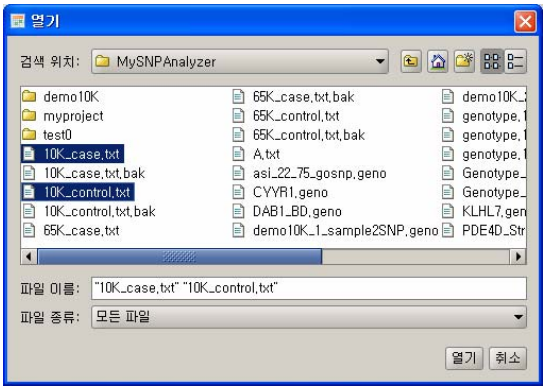
<Figure 2-5> Select genotype file

2.1.4. Import Genotype Data (Affymetrix GeneChip Data)

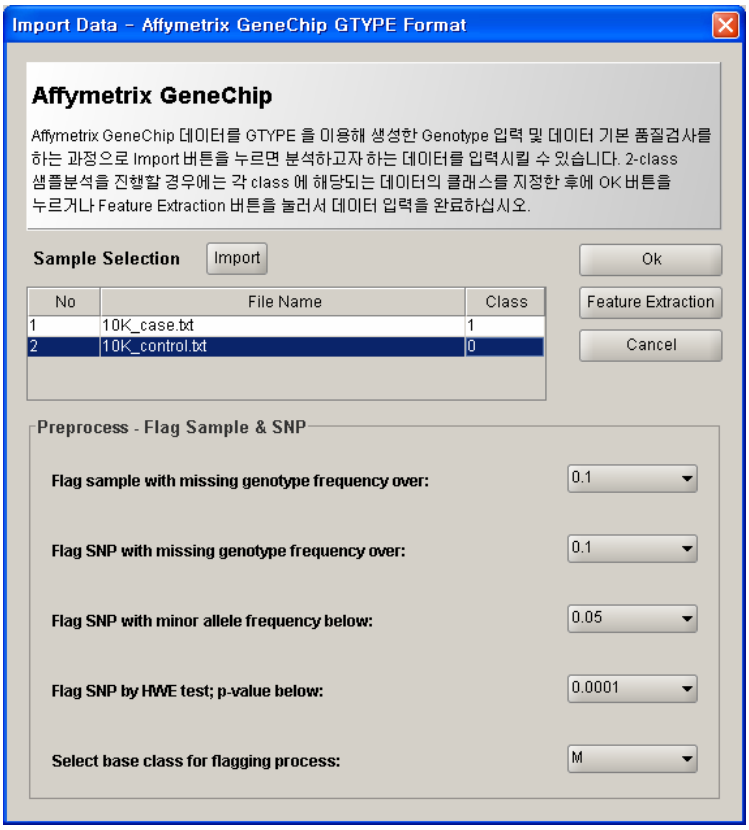
Click [File] > [Import Data] > [Affymetrix GeneChip GTYPE Format] in the main menu to display the window in which you can enter genotype data as in <Figure 2-6>. Click [Import] to display the window in which you can select genotype data as in <Figure 2-7>. Use the [CTRL] button to select up to 2 genotype files by class and click [Open] to display the selected genotypes in "Class Sample List" as in <Figure 2-8>. Use "Class" in the right of the list to select the sample type of each genotype. ("0" for control class and "1" for case class). File combinations that can be entered are the following:

- For 500K
  - One Control Class File
  - One Case Class File
- For 250K (NSP, STY Format)
  - When there exists only one of NSP or STY format.
    - One NSP (or STY) Control Class File
    - One NSP (or STY) Case Class File
  - When there exist both NSP and STY format data.
    - Each of NSP and STY Format Control Class File
    - Each of NSP and STY Format Case Class File

<Figure 2-6> Import Affymetrix GeneChip GTYPE format genotype data

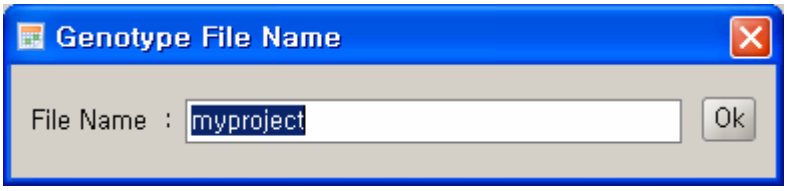


<Figure 2-7> Select genotype data

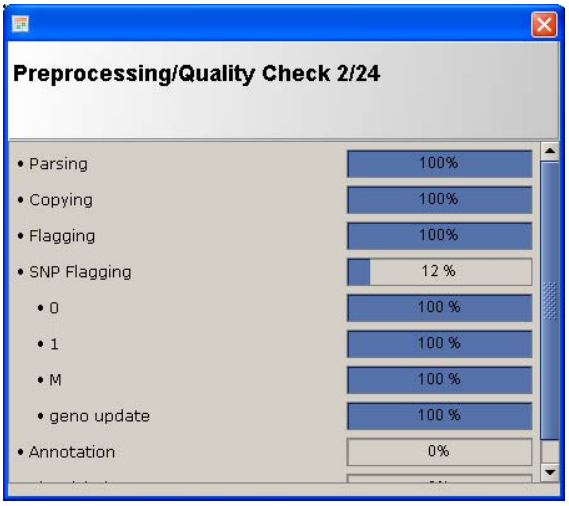


<Figure 2-8> Class setting of genotype

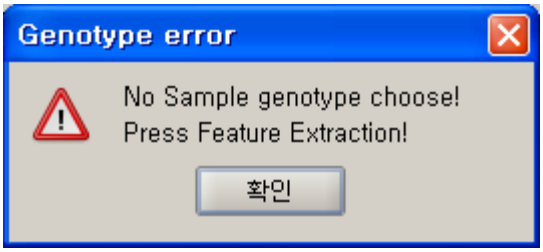
After you finish setting class, click [OK] in <Figure 2-8> and a window in which you can enter genotype file name to be used in the analysis process appears as in <Figure 2-9>. (If you perform Cross Tabulation Analysis using SNP, the result file name is like file name.chromosome number.snp.crss). Click [OK] and progress bar appears. If data format is not appropriate, a warning window shows as in <Figure 2-11>. (Please refer to **Chapter 7 Data Format** for more information). You can set parameters used for performing preprocess in “PreProcess – Flag Sample & SNP” after completing data input. (For more information on PreProcess, please refer to **Chapter 3 PreProcess.**)



<Figure 2-9> Input file name



<Figure 2-10> Genotype data input progress window



<Figure 2-11> > Data format error notification window

If the entered genotype format is not automatically recognizable, click [Feature Extraction] in <Figure 2-8> to show a window in which the user can specify the feature to extract from genotype as in <Figure 2-12>.

①: Select a sample class (control, case).

②: Items included in data (Feature Name) and each item's contents.

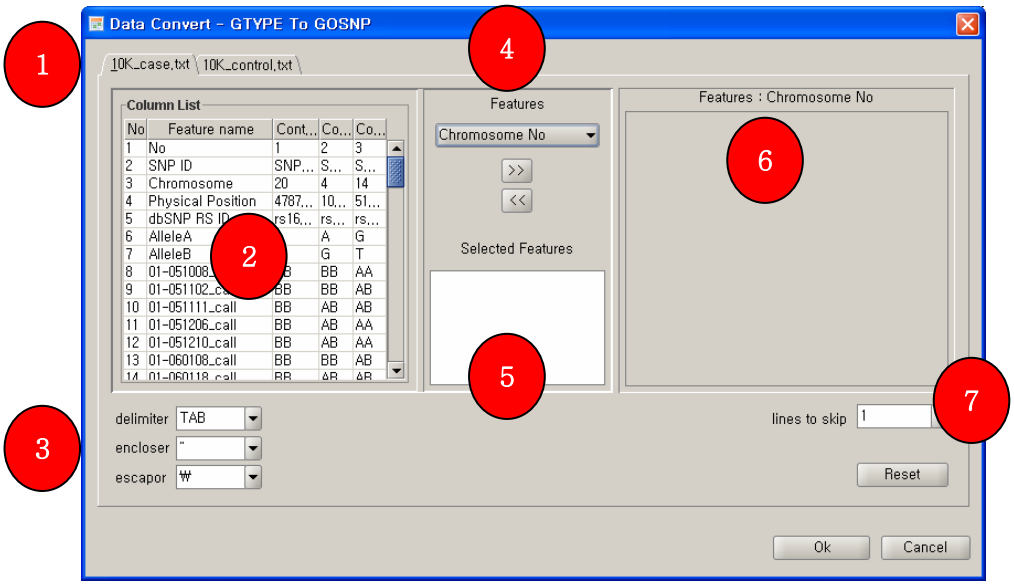
③: Identifier for identifying each item of data.

④: Required six items to extract from data. (Chromosome No, Allele, Marker ID, dbSNP #rs, Chromosome Position, Genotype)

⑤: Selected item list

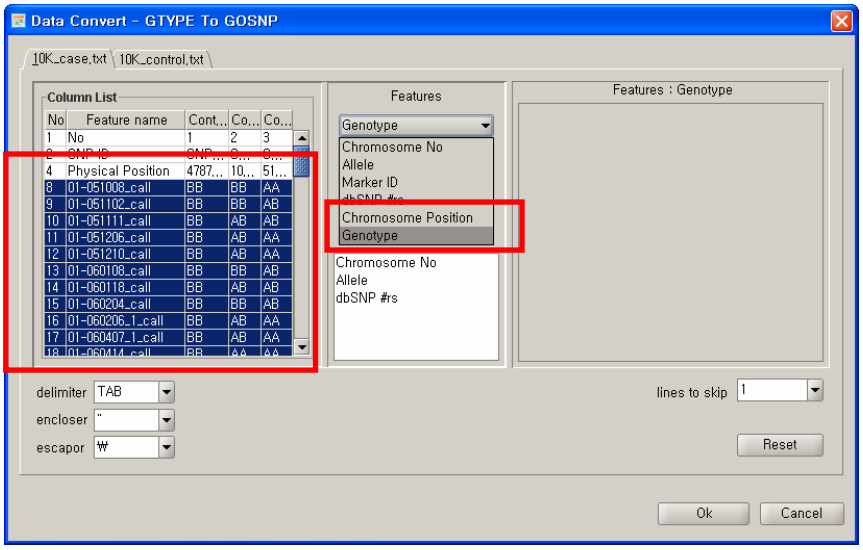
⑥: Selected item contents.

⑦: Number of lines to skip in the contents included in data (for header deletion)

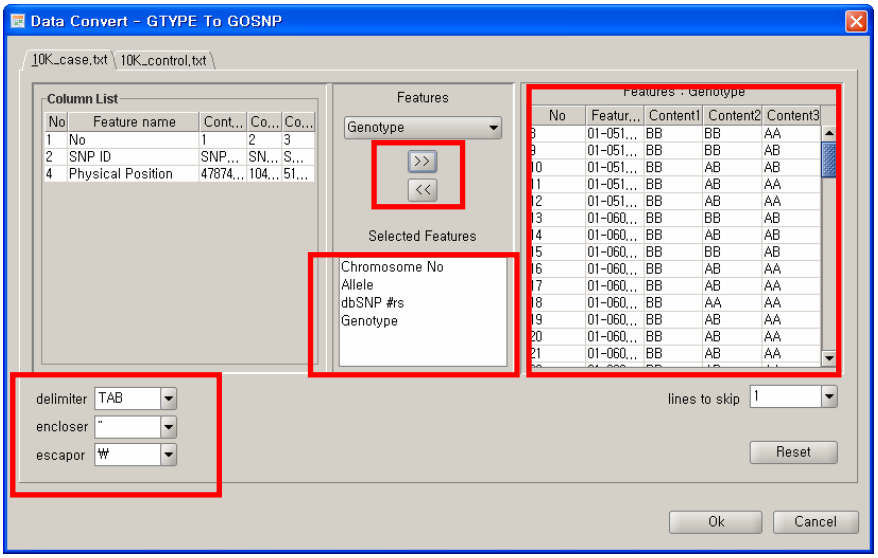


<Figure 2-12> Feature extraction interface

Set the item of "Feature" list to "Genotype" after selecting a sample genotype you want to extract from "Column List" as in <Figure 2-13>. Click [>>] and the specified item will be extracted as in <Figure 2-14>. You can extract the rest of items in the same manner. (Extract other class files in the same manner).

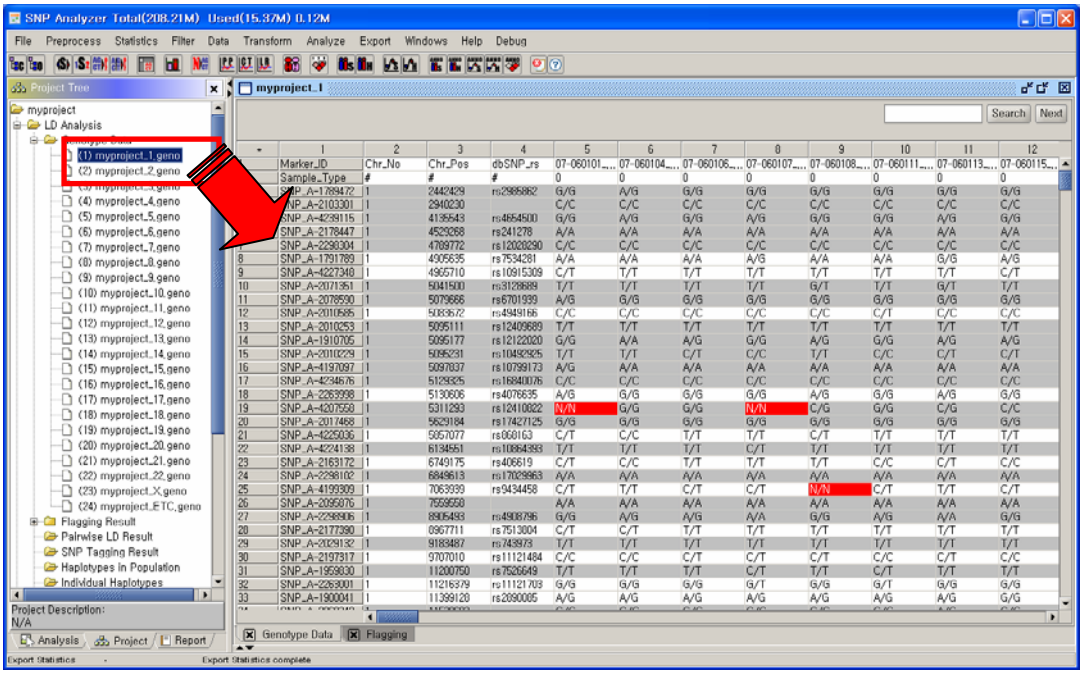


<Figure 2-13> Extract sample genotype



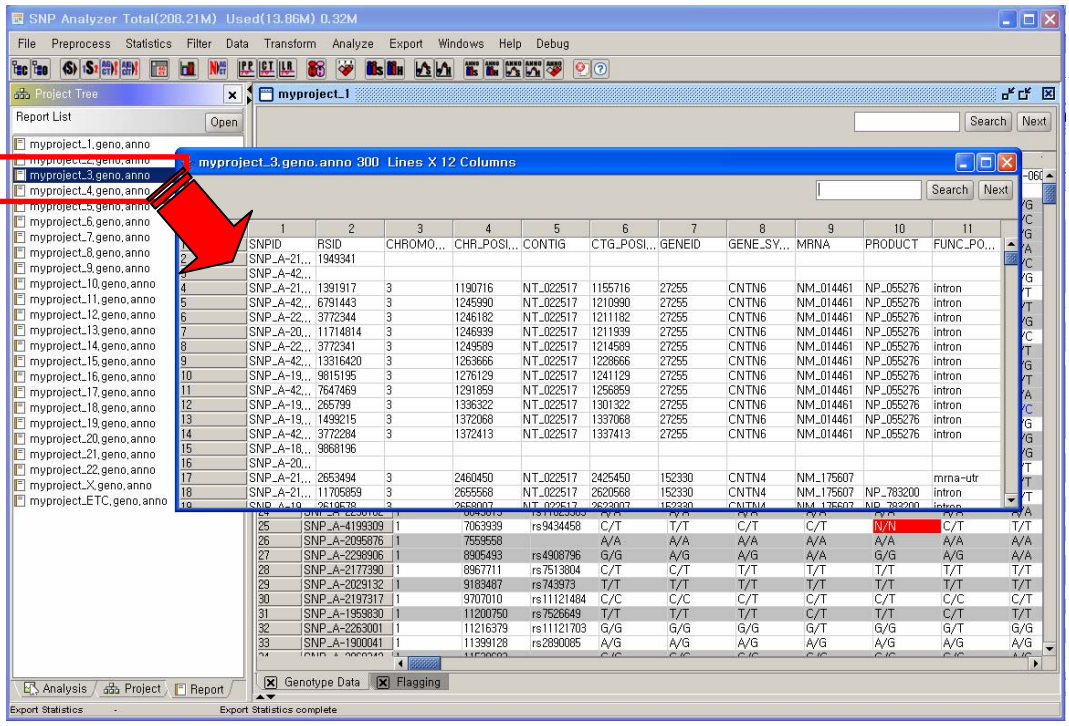
<Figure 2-14> Result of sample genotype extraction

Click [OK] after setting items to extract. If you set items to extract appropriately, progressive bar in <Figure 2-10> appears and a warning window in <Figure 2-11> shows otherwise. After completing all data input progress, input data are placed in project tree as in <Figure 2-15>. (Genotype data are sorted by chromosome number and saved individually). Select and double-click each of genotype data in project tree to display the content of the corresponding genotype on the main screen: missing genotype in red and excluded SNPs in gray.



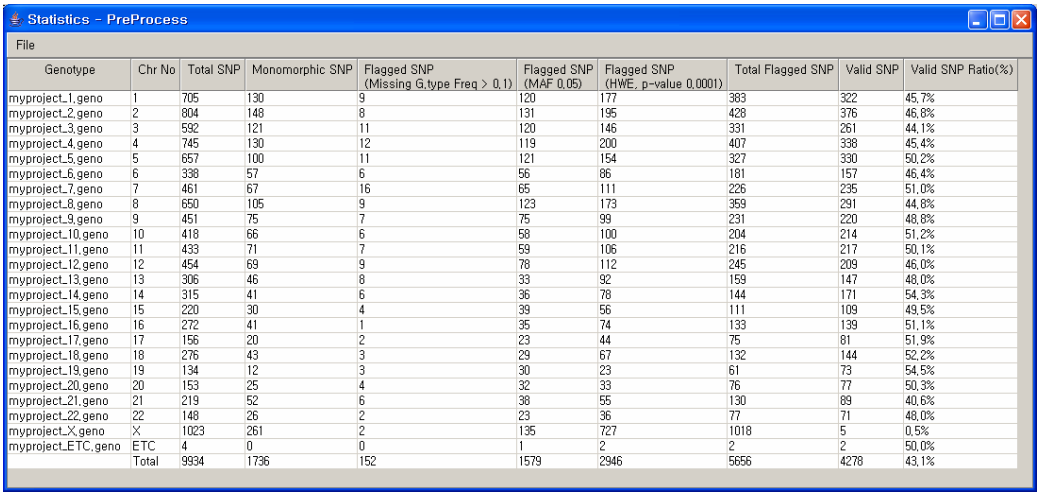
<Figure 2-15> Genotype data by chromosome number in project data

<Figure 2-16> shows the contents of annotation file when double-clicking one of the annotation file lists created in the process of data input. (For annotation file formats, please refer to **Chapter 7 Data Format**).



<Figure 2-16> Annotation file

After completing the data input process, the statistics of the data preprocess result shows as a pop-up window as in <Figure 2-17>. For the details of each item, please refer to **Chapter 3 PreProcess**.

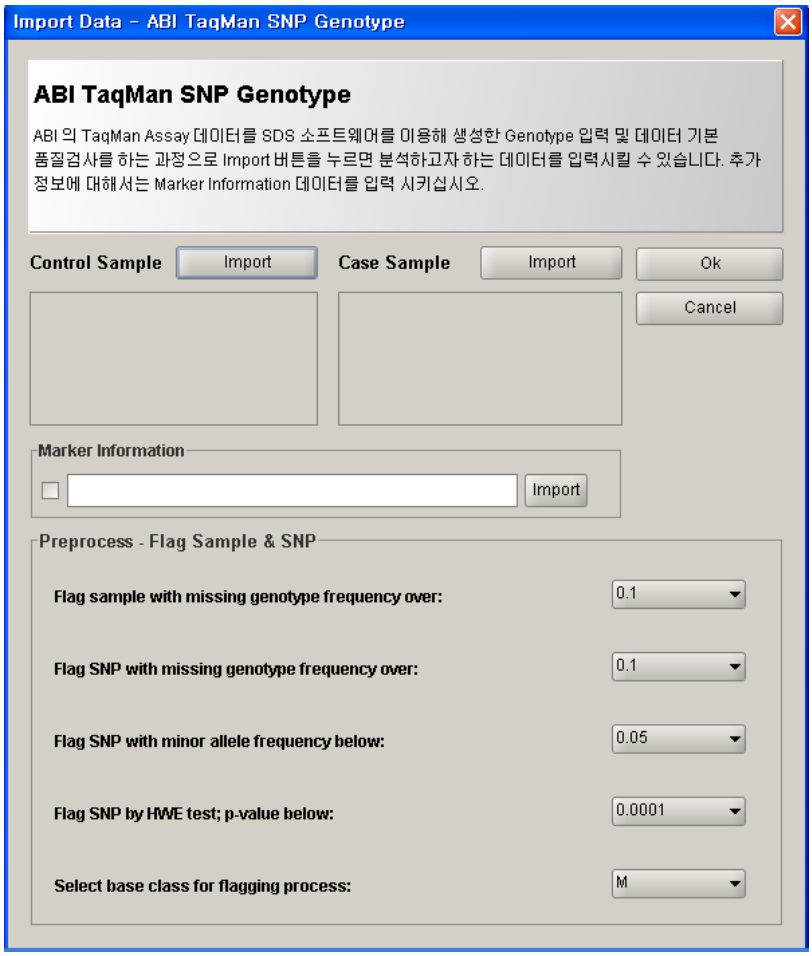


Genotype	Chr No	Total SNP	Monomorphic SNP	Flagged SNP (Missing G.type Freq > 0.1)	Flagged SNP (MAF 0.05)	Flagged SNP (HWE, p-value 0.0001)	Total Flagged SNP	Valid SNP	Valid SNP Ratio(%)
myproject_1.geno	1	705	130	9	120	177	383	322	45.7%
myproject_2.geno	2	804	148	8	131	195	428	376	46.8%
myproject_3.geno	3	592	121	11	120	146	331	261	44.1%
myproject_4.geno	4	745	130	12	119	200	407	338	45.4%
myproject_5.geno	5	657	100	11	121	154	327	330	50.2%
myproject_6.geno	6	338	57	6	56	86	181	157	46.4%
myproject_7.geno	7	461	67	16	65	111	226	235	51.0%
myproject_8.geno	8	650	105	9	123	173	359	291	44.8%
myproject_9.geno	9	451	75	7	75	99	231	220	48.8%
myproject_10.geno	10	418	66	6	58	100	204	214	51.2%
myproject_11.geno	11	433	71	7	59	106	216	217	50.1%
myproject_12.geno	12	454	69	9	78	112	245	209	46.0%
myproject_13.geno	13	306	46	8	33	92	159	147	48.0%
myproject_14.geno	14	315	41	6	36	78	144	171	54.3%
myproject_15.geno	15	220	30	4	39	56	111	109	49.5%
myproject_16.geno	16	272	41	1	35	74	133	139	51.1%
myproject_17.geno	17	156	20	2	23	44	75	81	51.9%
myproject_18.geno	18	276	43	3	29	67	132	144	52.2%
myproject_19.geno	19	134	12	3	30	23	61	73	54.5%
myproject_20.geno	20	153	25	4	32	33	76	77	50.3%
myproject_21.geno	21	219	52	6	38	55	130	89	40.6%
myproject_22.geno	22	148	26	2	23	36	77	71	48.0%
myproject_X.geno	X	1023	261	2	135	727	1018	5	0.5%
myproject_ETC.geno	ETC	4	0	0	1	2	2	2	50.0%
	Total	9934	1736	152	1579	2946	5656	4278	43.1%

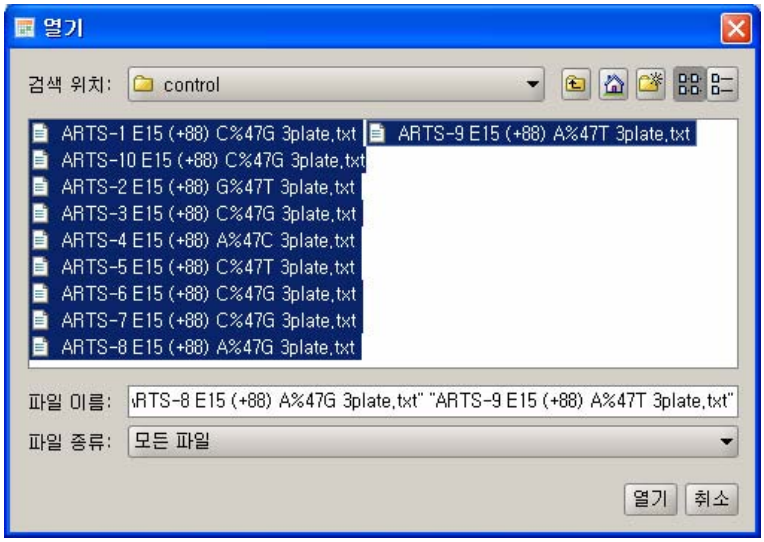
<Figure 2-17> Statistics for input and preprocess data

2.1.5. Import Genotype Data (ABI TaqMan Data)

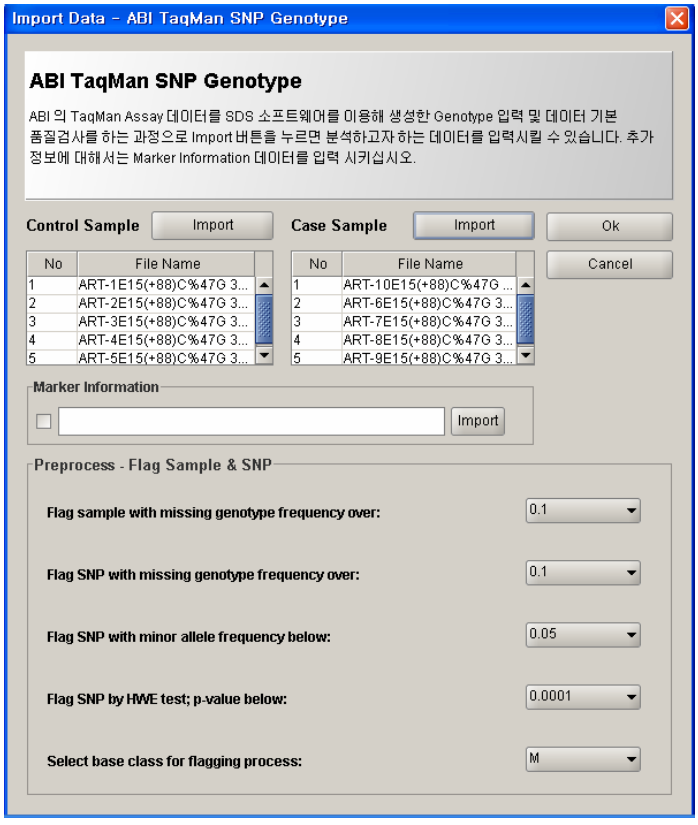
Click [File] > [Import Data] > [ABI TaqMan SNP Genotype] in the main menu to show the window where you can enter genotype data as in <Figure 2-18>. Click [Import] in "Control Sample" to show the window where you can select genotype files as in <Figure 2-19>. Select files using the [CTRL] or [SHIFT] key to input files. Case sample genotype file can be entered in the same manner as control sample by clicking [Import] in "Case Sample". <Figure 2-20> shows the result of file input. If there is a marker information file for SNP, click [Import] of "Marker Information" and enter the corresponding file.



<Figure 2-18> Import ABI TaqMan genotype data



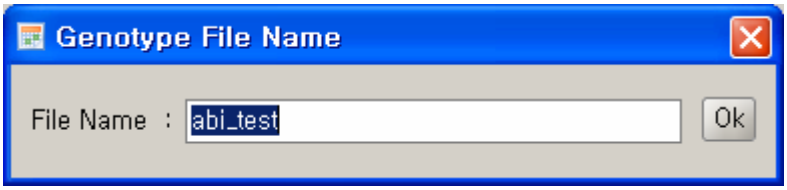
<Figure 2-19> Select genotype data of control sample



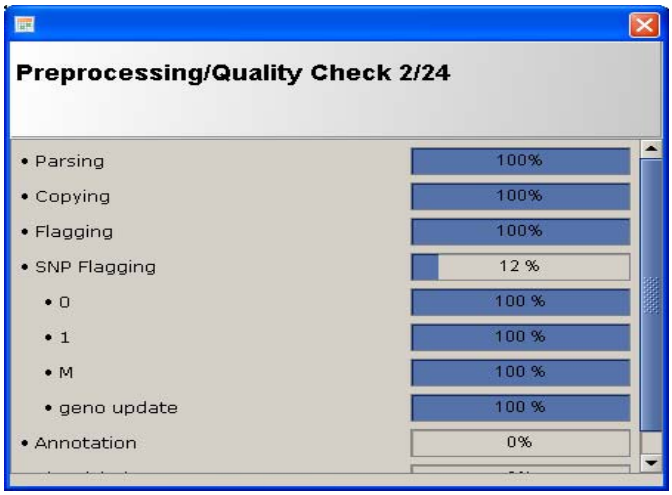
<Figure 2-20> Completion of genotype data input

Click [OK] after completing genotype file input to show the window where you can input a genotype file name to use for analysis as in <Figure 2-21>. (If Cross Tabulation Analysis is performed using SNP, the result file name is like filename.chromosome number.snp.crss). Click

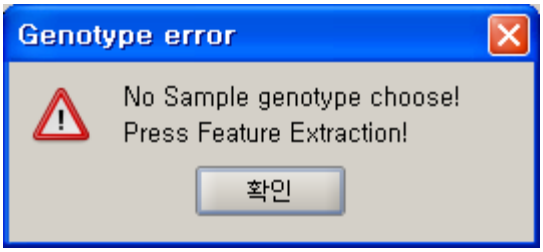
[OK] and progressive bar appears as in <Figure 2-22>. If data format is not appropriate, a warning window appears as in <Figure 2-23>. (Please refer to **Chapter 7 Data Format** for the details.) You can set parameters implemented during preprocess in “PreProcess – Flag Sample & SNP” after completiing data input. (Please refer to **Chapter 3 PreProcess** for the details on preprocess).



<Figure 2-21> Input file name

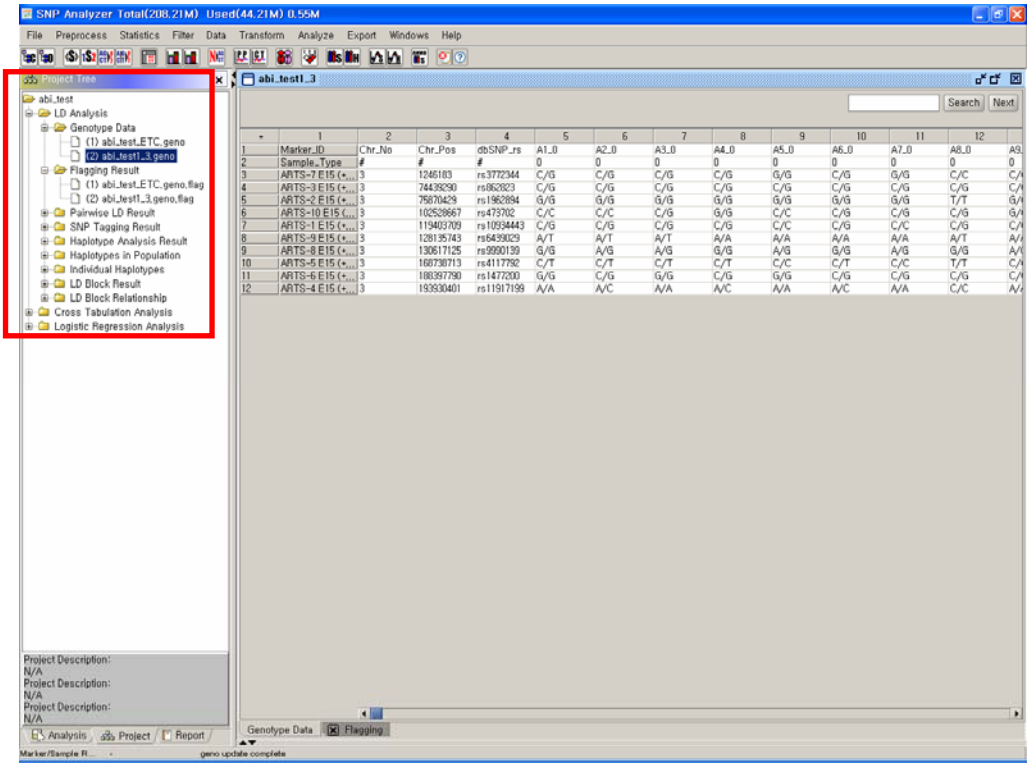


<Figure 2-22> Genotype data input progress window



<Figure 2-23> > Data format error notification window

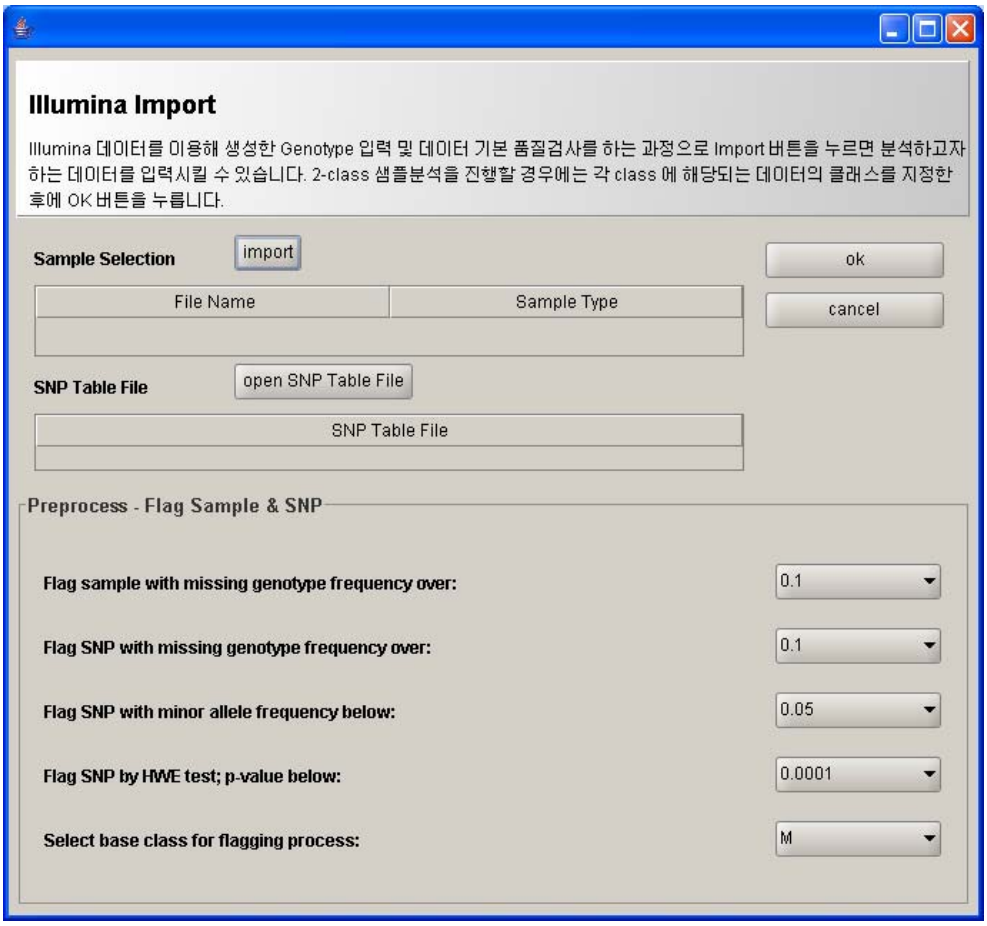
When all data input progress is completed, input data are placed in project tree as in <Figure 2-23>.



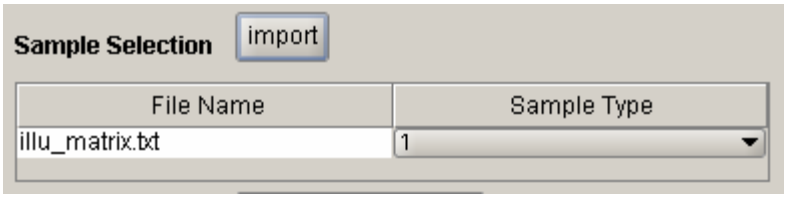
<Figure 2-24> Genotype data in project tree

2.1.6. Import Genotype Data (Illumina Data)

Click [File] > [Import Data] > [Illumina Matrix Format Import] in the main menu and a window where you can input genotype data appears as in <Figure 2-25>. Click [Import] and a window where you can select genotype data appears as in <Figure 2-5>. Use the [CTRL] button to select genotype files by class and click [Open] to display the selected genotype in "Sample Selection" as in <Figure 2-26>. Select a sample type for each genotype using "Sample Type" on the right side of the list. ("0" for control sample and "1" for case sample.) The file combinations you can input are the following:



<Figure 2-25> Illumina matrix format import



<Figure 2-26> Sample type setting

After completing class setting, click [OK] and progressive bar appears as in <Figure 2-10>.

When all data input progress is completed, input data are placed in project tree as in <Figure 2-15>. (Genotype data are sorted by chromosome number and saved as individual files.) Select and double-click one of genotype data in project tree and the content of the corresponding genotype is displayed in the main screen: missing genotype in red and excluded SNPs in gray.

After completing data input process, statistics for data preprocess result shows as a pop-up as in <Figure 2-17>. For the details on each item, please refer to **Chapter 3 PreProcess.**

Chapter 3

PreProcess

3. PreProcess

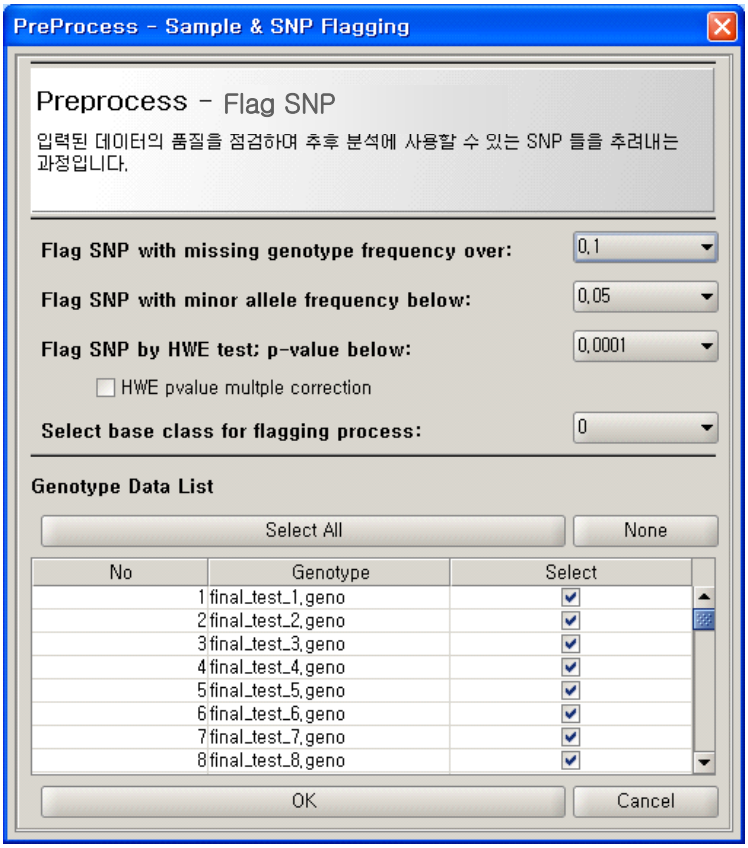
It removes SNPs and samples to be excluded for further analysis or replaces missing genotypes with other observed genotypes.

3.1. Flag Sample & SNP

If the quality of input data is not good enough, you can set preprocess parameters differently by file considering the property of input data. Click [PreProcess] > [Flag SNP] to set parameters as in <Figure 3-1>.

- Flag SNP with missing genotype frequency over: remove the corresponding SNP if missing genotype observed is bigger than the set value (default = 0.5).
- Flag SNP with minor allele frequency below: remove the corresponding SNP if minor allele frequency observed is smaller than the set value (default = 0.05).
- Flag SNP by HWE test; p-value below: remove the corresponding SNP if the calculated p-value is smaller than the set value after Hardy-Weinberg Equilibrium (default = 0.0001).
- Select base class for flagging process: set the base sample to perform the HWE test.

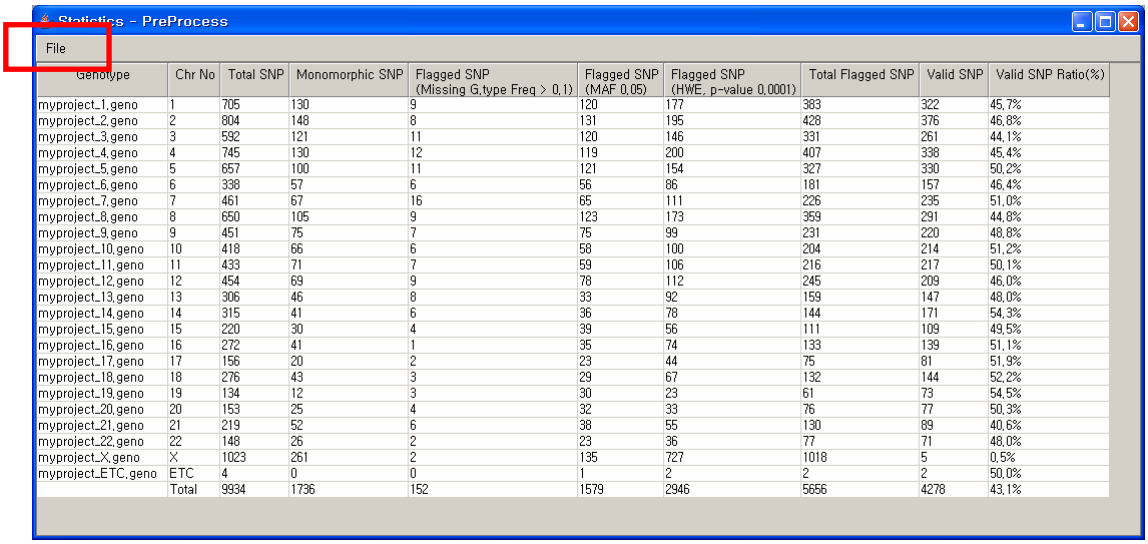
Click [Select All] and click [OK] in <Figure 3-1> to perform preprocess on all the input genotypes.



<Figure 3-1> Data preprocess option setting

The results are added in project tree after completing preprocess and the statistical result is displayed in table as in <Figure 3-2>. Click [File] > [Save] to save the corresponding statistical result and the saved result is added in "Report" of project tree. Description for each item of the table is the following:

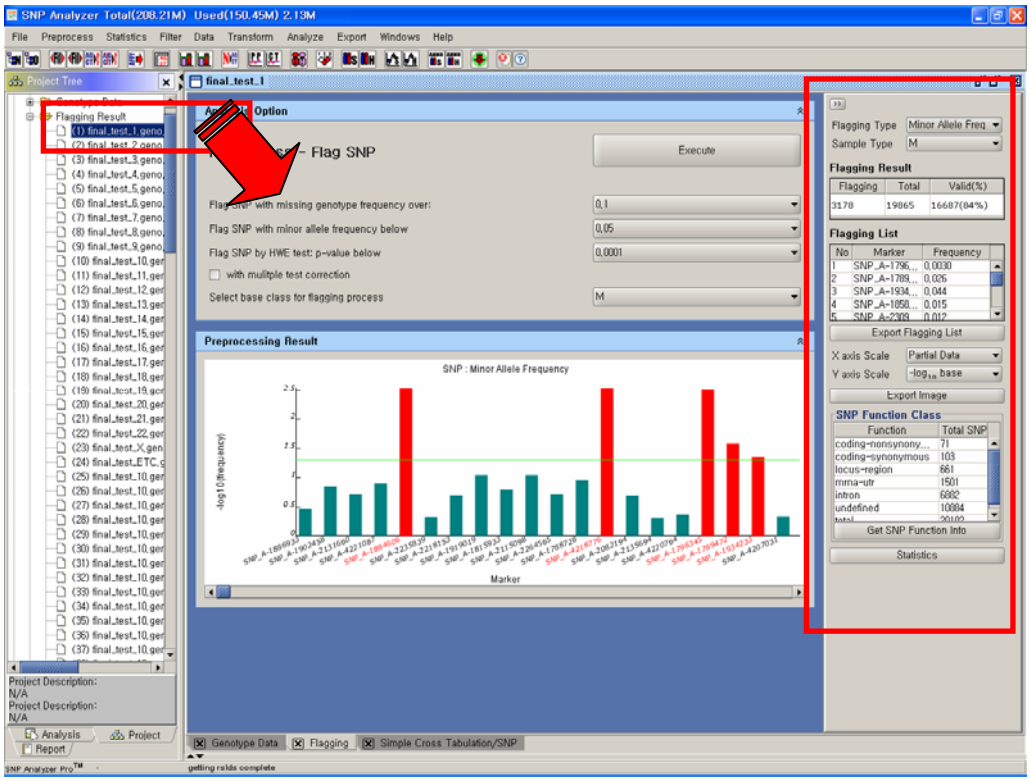
- Genotype: Genotype file names for preprocess
- Chr No: Chromosome number of the specified genotype file
- Total SNP: Total number of SNPs in a specified genotype before preprocess
- Monomorphic SNP: Number of SNPs with only one genotype
- Flagged SNP (Missing G.Type Freq > 0.5): Number of removed SNPs with missing genotype frequency over 0.5
- Flagged SNP (MAF < 0.05): Number of removed SNPs with minor allele frequency less than 0.05
- Flagged SNP (HWE, p-value < 0.0001): Number of removed SNPs with p-value less than 0.0001 after HWE test
- Valid SNP: Number of SNPs passing the preprocess
- Valid SNP Ratio (%): (Valid SNP Number / Total SNP Number) x 100 (%)



Genotype	Chr No	Total SNP	Monomorphic SNP	Flagged SNP (Missing G,type Freq > 0,1)	Flagged SNP (MAF 0,05)	Flagged SNP (HWE, p-value 0,0001)	Total Flagged SNP	Valid SNP	Valid SNP Ratio(%)
myproject_1.geno	1	705	130	9	120	177	383	322	45,7%
myproject_2.geno	2	804	148	8	131	195	428	376	46,8%
myproject_3.geno	3	592	121	11	120	146	331	261	44,1%
myproject_4.geno	4	745	130	12	119	200	407	338	45,4%
myproject_5.geno	5	657	100	11	121	154	327	330	50,2%
myproject_6.geno	6	338	57	6	56	86	181	157	46,4%
myproject_7.geno	7	461	67	16	65	111	226	235	51,0%
myproject_8.geno	8	650	105	9	123	173	359	291	44,8%
myproject_9.geno	9	451	75	7	75	99	231	220	48,8%
myproject_10.geno	10	418	66	6	58	100	204	214	51,2%
myproject_11.geno	11	433	71	7	59	106	216	217	50,1%
myproject_12.geno	12	454	69	9	78	112	245	209	46,0%
myproject_13.geno	13	306	46	8	33	92	159	147	48,0%
myproject_14.geno	14	315	41	6	36	78	144	171	54,3%
myproject_15.geno	15	220	30	4	39	56	111	109	49,5%
myproject_16.geno	16	272	41	1	35	74	133	139	51,1%
myproject_17.geno	17	156	20	2	23	44	75	81	51,9%
myproject_18.geno	18	276	43	3	29	67	132	144	52,2%
myproject_19.geno	19	134	12	3	30	23	61	73	54,5%
myproject_20.geno	20	153	25	4	32	33	76	77	50,3%
myproject_21.geno	21	219	52	6	38	55	130	89	40,6%
myproject_22.geno	22	148	26	2	23	36	77	71	48,0%
myproject_X.geno	X	1023	261	2	135	727	1018	5	0,5%
myproject_ETC.geno	ETC	4	0	0	1	2	2	2	50,0%
	Total	9934	1736	152	1579	2946	5656	4278	43,1%

<Figure 3-2> Statistics for input and preprocess data

To view the details of preprocess result, select and double-click one of the preprocess result data after selecting the "Project" tab under the project tree.

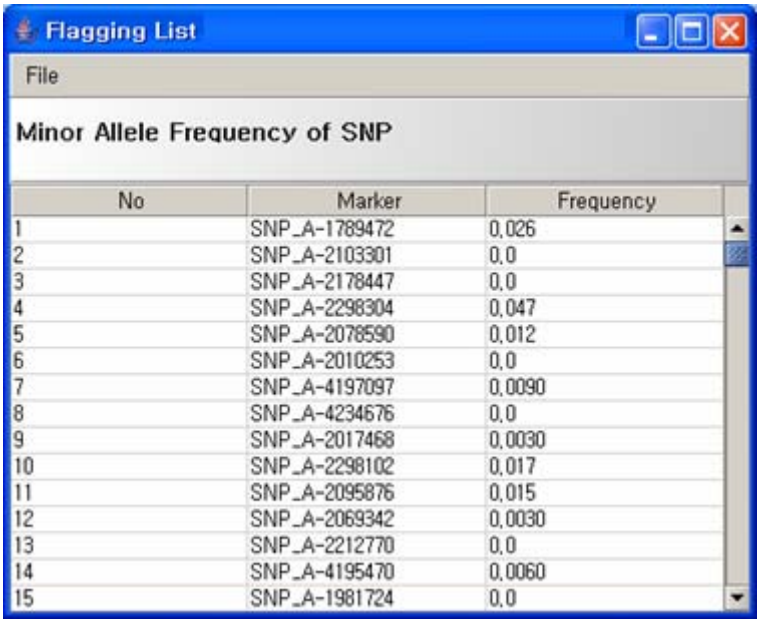


<Figure 3-3> Result of preprocess

3.1.1. Graph Visualization and Result Saving Panel

You can control the visualization format in the right panel of <Figure 3-3>. The details are the following:

- Flagging Type: Select one of Flagging methods
  - Sample – Call Rate
  - Missing Genotype Frequency
  - Minor Allele Frequency
  - HWE Test
- Sample Type: Select one of Input Samples
  - 0: Control Sample
  - 1: Case Sample
  - M: Result of Integrating Control Sample and Case Sample
- Flagging Result
  - Flagging: Number of removed SNPs in the selected preprocess method
  - Total: Total number of SNPs in the selected file
  - Valid (%): (Number of remaining SNPs after preprocess/Total SNP number) *100
- Flagging List: List of removed SNPs by preprocess
  - Click [Export Flagging List] to display information of the removed SNPs by preprocess as in <Figure 3-4>. Click [File] > [Save] to save the result in text file.



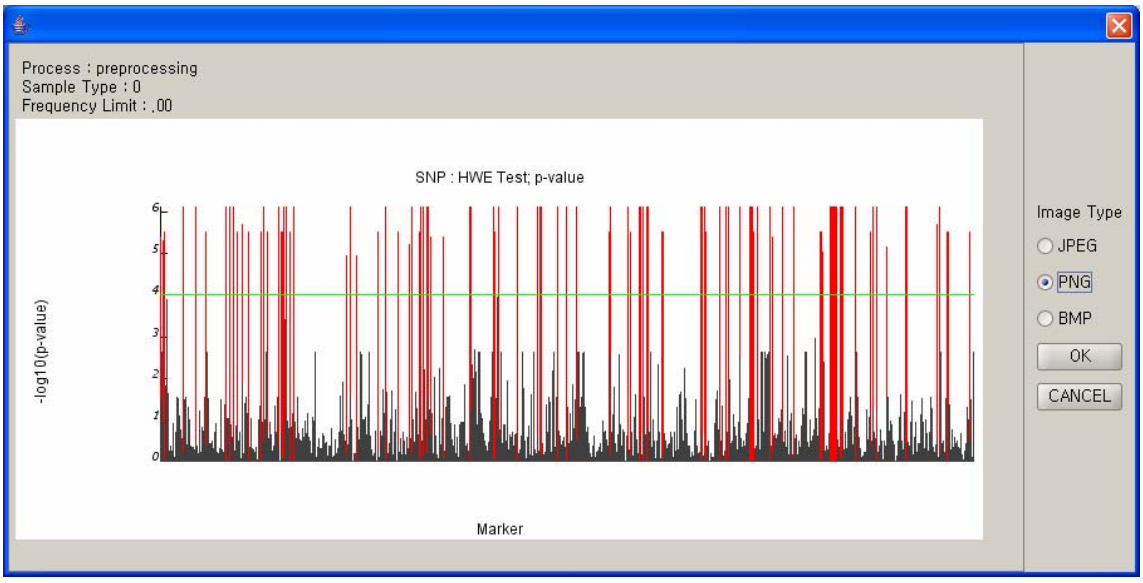
No	Marker	Frequency
1	SNP_A-1789472	0,026
2	SNP_A-2103301	0,0
3	SNP_A-2178447	0,0
4	SNP_A-2298304	0,047
5	SNP_A-2078590	0,012
6	SNP_A-2010253	0,0
7	SNP_A-4197097	0,0090
8	SNP_A-4234676	0,0
9	SNP_A-2017468	0,0030
10	SNP_A-2298102	0,017
11	SNP_A-2095876	0,015
12	SNP_A-2069342	0,0030
13	SNP_A-2212770	0,0
14	SNP_A-4195470	0,0060
15	SNP_A-1981724	0,0

<Figure 3-4> List of removed SNPs by preprocess

- X axis Scale: Set partial or whole number of SNPs for visualization
  - Partial Data: visualize the result of 20 SNPs

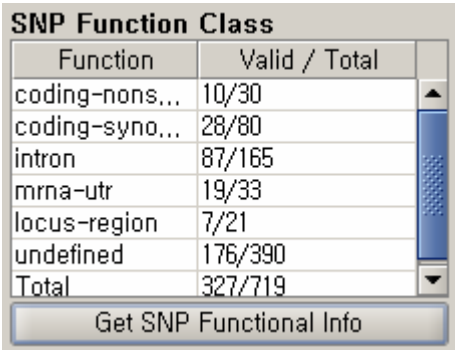
- Whole Data: visualize the result of whole SNPs

- Y Axis Scale: Set the unit for Y axis
  - -log10 base: show as -log10 (actual value)
  - -log2 base: show as -log2 (actual value)
  - Frequency: show as actual values
- Click [Export Image] to save the result as figure file of JPEG, PNG, and BMP formats. Click the "Report" tab under the project tree to view the saved figure files.



<Figure 3-5> List of removed SNPs by preprocess

- SNP Function Class: functional class information of the SNPs
  - Click [Get SNP Functional Info] to display functional class of the SNPs as in <Figure 3-6>.
  - Function: Defined by dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
    - Coding-nonsynonymous
    - Coding-synonymous
    - Intron
    - Mrna-utr
    - Locus-region
    - Undefined: Without locus information
  - Valid/Total: Number of function of SNPs after preprocess/ Number of total function of SNPs

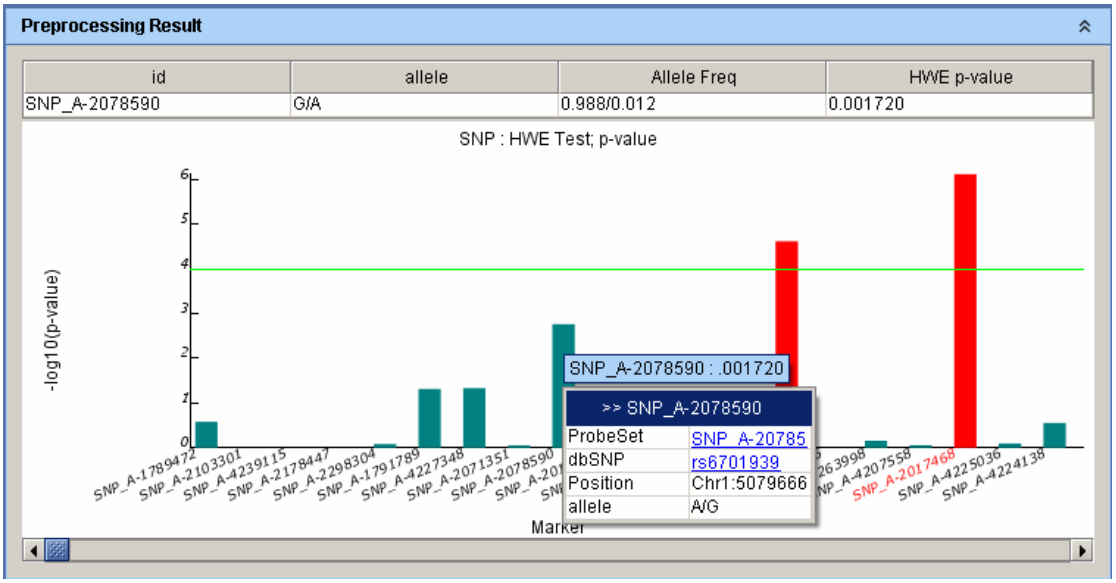


<Figure 3-6> SNP function class information

- Click [Statistics] to show the statistical results in the same way as in <Figure 3-2>.

3.1.2. PreProcess Control and Graph Panel

The upper middle screen in <Figure 3-3> is the panel you can perform preprocess again for SNPs that are currently displayed. Click [Execute] to perform preprocess after setting parameters. Preprocess results are shown in bar chart and the green horizontal line in graph means the threshold. <Figure 3-7> is an example of result graph. SNPs displayed in red are the removed SNPs. If you right-click the specified SNP, information about the SNP is shown in a pop-up window. Click dbSNP #rs number to connect to dbSNP site and view detailed information of the specified SNP.



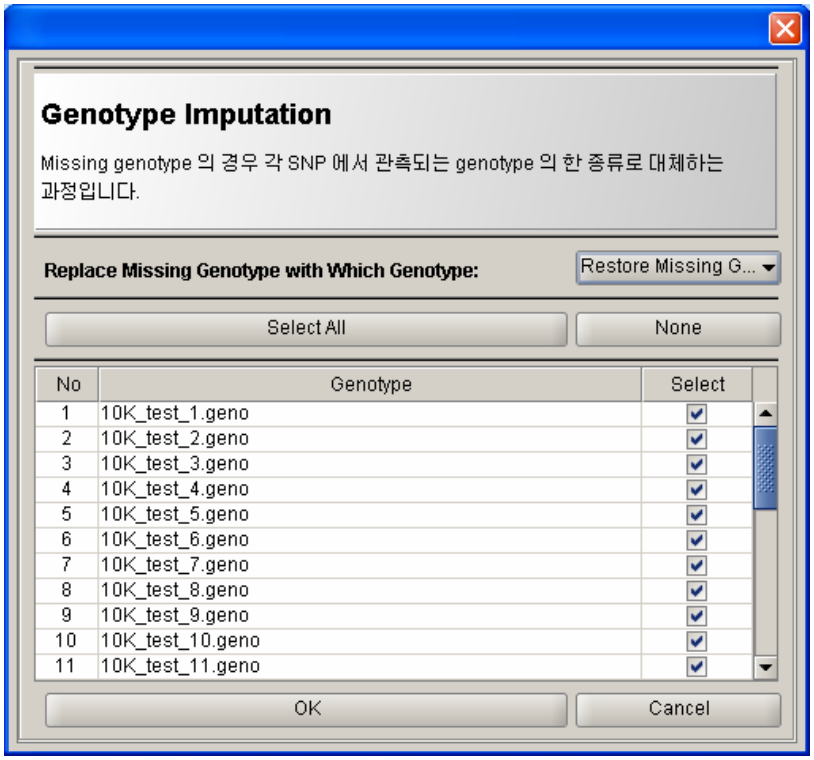
<Figure 3-7> Preprocess result graph

3.1.3. Replace Missing Genotype

Click [PreProcess] > [Replace Missing Genotype], and a window where you can replace each

SNP with one of observed genotypes appears. Replaced genotypes are the following:

- Restore missing genotype: restore replaced genotype back to original missing genotype
- Hetero genotype: replace with heterozygous genotype observed in a specified SNP
- Major homo genotype: replace with major homozygous genotype observed in a specified SNP
- Minor homo genotype: replace with minor homozygous genotype observed in a specified SNP



<Figure 3-8> Missing genotype imputation

Chapter 4

Analyze

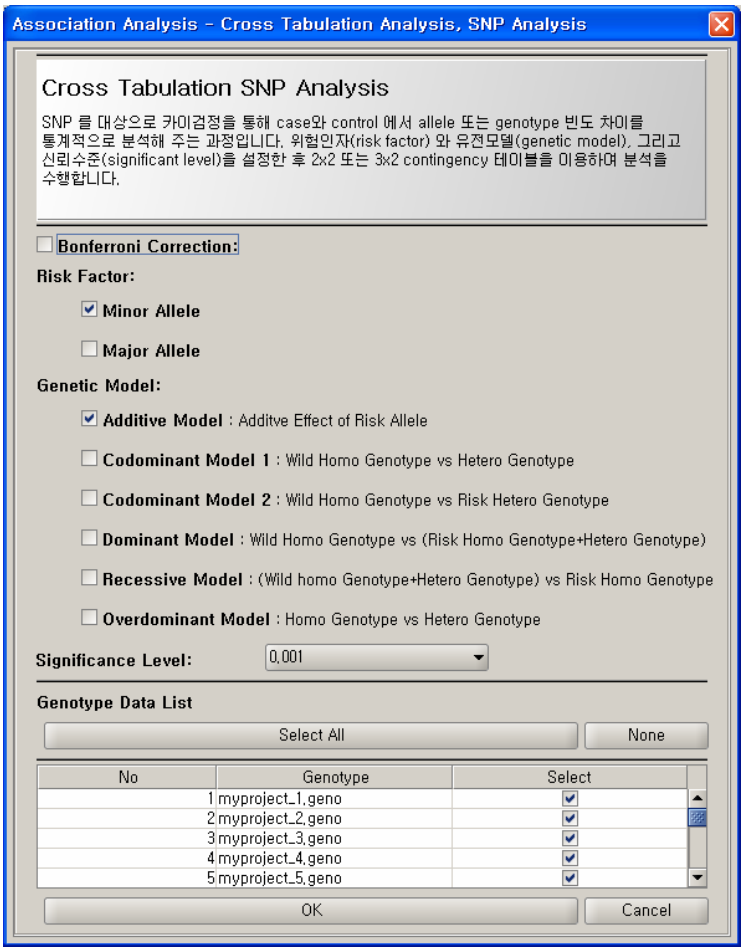
4. Association Analysis/LD Analysis

Cross tabulation analysis in the association analysis menu performs chi-square test to extract SNPs or haplotypes of which allele/genotype frequencies or haplotype frequencies are significantly different in case sample and control sample. Logistic regression analysis extracts SNPs and haplotypse that can well discriminate between case sample and control sample using binary logistic regression model. LD analysis estimates linkage disequilibrium between SNPs and reconstructs LD blocks.

4.1. Cross Tabulation Analysis using SNP

Click [Analyze] > [Association Analysis] > [Cross Tabulation Analysis] > [SNP Analysis] to show a window like <Figure 4-1>. You can perform many kinds of analysis models multiply by considering risk factor. Please refer to **Appendix-B** for detailed information about analysis model and estimated values.

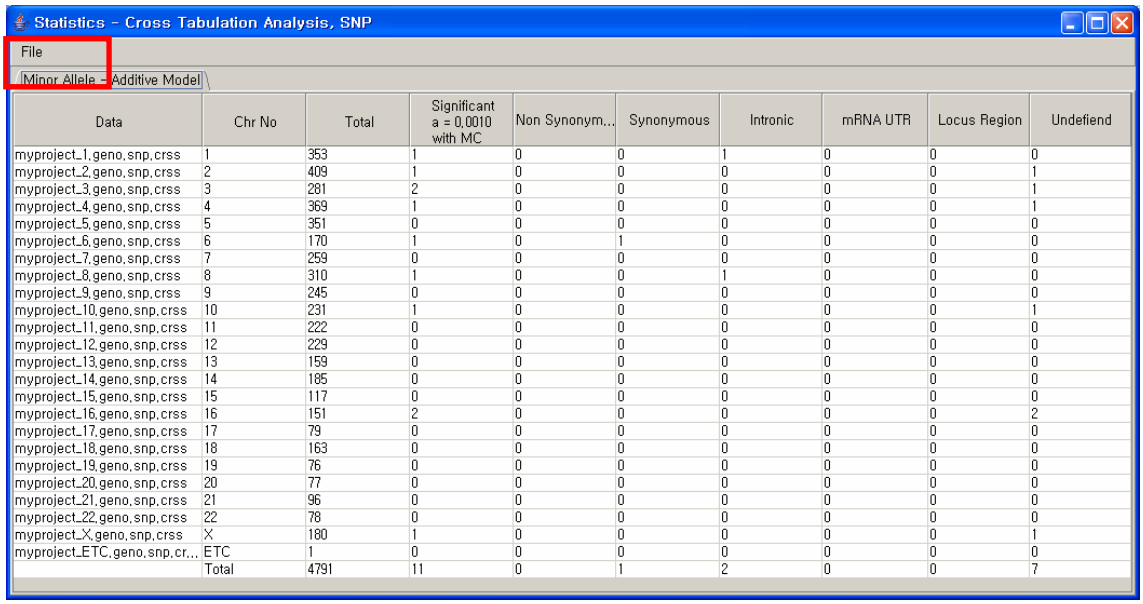
- Risk Factor
  - Minor Allele: minor allele observed in each SNP
  - Major Allele: major allele observed in each SNP
- Genetic Model
  - Additive Model
  - Codominant Model 1
  - Codominant Model 2
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Estimated Value
  - Odds Ratio
  - Attributable Risk %
  - Population Attributable Risk %



<Figure 4-1> Cross Tabulation Analysis setting window

Select genotype files from the “Genotype Data List” after setting risk factor, genetic model and the significance level. The analysis result is added in project tree after completing analysis and the statistic result appears as a pop-up window like <Figure 4-2>. Click [File] > [Save] to save the specified statistics and the saved result is added in "Report" tab of project tree. Descriptions for each item in statistic table are the following:

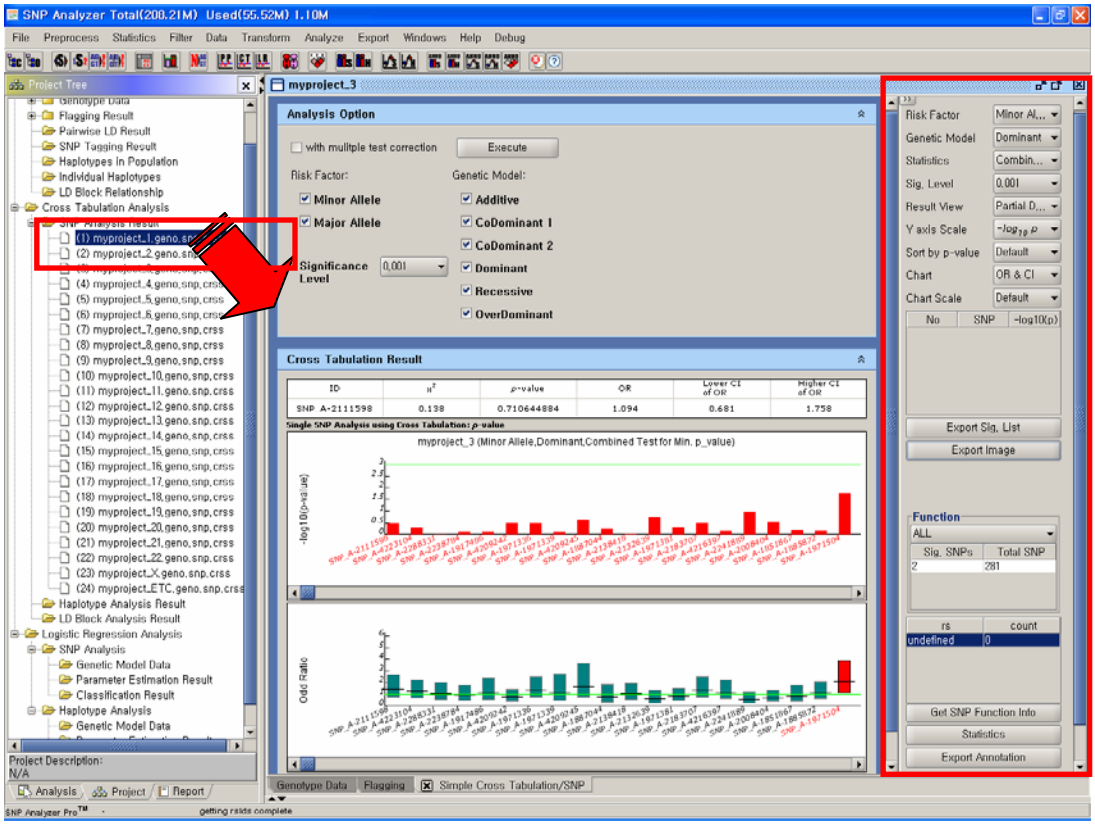
- Data: File list for the analysis
- Chr No: Chromosome number of a specified file
- Total: Total number of SNPs in a specified file
- Significant (a=0.001 with MC): Number of significant SNPs (significance level a=0.001) with multiple test correction (Bonferroni Correction)
- Function Class: Non Synonymous, Synonymous, Intron, mRNA UTR, Locus Region, Undefined



Data	Chr No	Total	Significant a = 0.0010 with MC	Non Synonym...	Synonymous	Intronic	mRNA UTR	Locus Region	Undefiend
myproject_1.geno.snp.crss	1	353	1	0	0	1	0	0	0
myproject_2.geno.snp.crss	2	409	1	0	0	0	0	0	1
myproject_3.geno.snp.crss	3	281	2	0	0	0	0	0	1
myproject_4.geno.snp.crss	4	369	1	0	0	0	0	0	1
myproject_5.geno.snp.crss	5	351	0	0	0	0	0	0	0
myproject_6.geno.snp.crss	6	170	1	0	1	0	0	0	0
myproject_7.geno.snp.crss	7	259	0	0	0	0	0	0	0
myproject_8.geno.snp.crss	8	310	1	0	0	1	0	0	0
myproject_9.geno.snp.crss	9	245	0	0	0	0	0	0	0
myproject_10.geno.snp.crss	10	231	1	0	0	0	0	0	1
myproject_11.geno.snp.crss	11	222	0	0	0	0	0	0	0
myproject_12.geno.snp.crss	12	229	0	0	0	0	0	0	0
myproject_13.geno.snp.crss	13	159	0	0	0	0	0	0	0
myproject_14.geno.snp.crss	14	185	0	0	0	0	0	0	0
myproject_15.geno.snp.crss	15	117	0	0	0	0	0	0	0
myproject_16.geno.snp.crss	16	151	2	0	0	0	0	0	2
myproject_17.geno.snp.crss	17	79	0	0	0	0	0	0	0
myproject_18.geno.snp.crss	18	163	0	0	0	0	0	0	0
myproject_19.geno.snp.crss	19	76	0	0	0	0	0	0	0
myproject_20.geno.snp.crss	20	77	0	0	0	0	0	0	0
myproject_21.geno.snp.crss	21	96	0	0	0	0	0	0	0
myproject_22.geno.snp.crss	22	78	0	0	0	0	0	0	0
myproject_X.geno.snp.crss	X	180	1	0	0	0	0	0	1
myproject_ETC.geno.snp.cr...	ETC	1	0	0	0	0	0	0	0
	Total	4791	11	0	1	2	0	0	7

<Figure 4-2> Cross Tabulation Analysis statistic result

Select and double-click one of the analysis results added in project tree and the analysis result is shown in graph as in <Figure 4-3>.



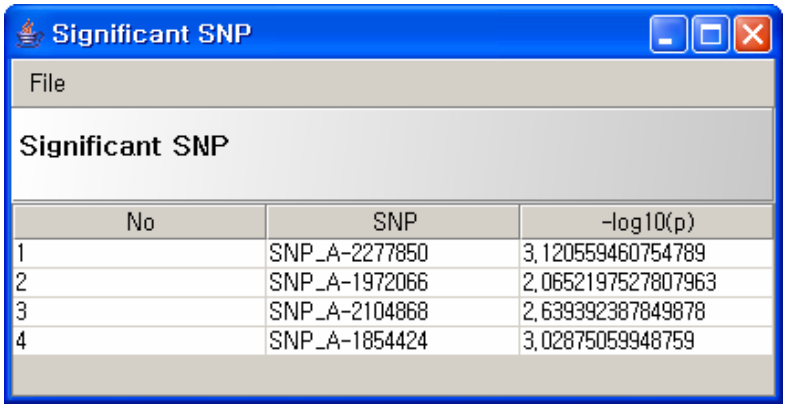
<Figure 4-3> Cross Tabulation Analysis result

4.1.1. Graph Visualization Control and Result Saving Panel

You can control many kinds of visualization parameters in the right panel of <Figure 4-3>. The details are the following:

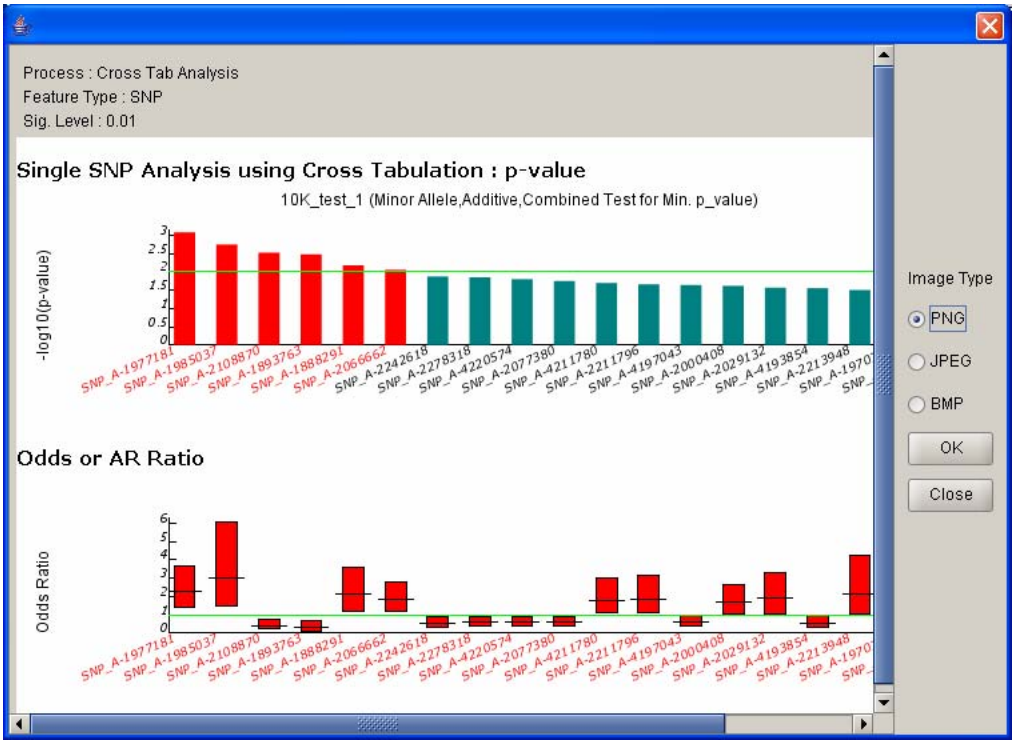
- Risk Factor: specify risk allele
  - Minor Allele
  - Major Allele
- Genetic Model: specify genetic model used in analysis
  - Additive Model
  - Codominant Model 1
  - Codominant Model 2
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Sig. Level: specify significance level
  - Setting Values: 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00001, 0.000001
- X axis Scale: specify number of SNPs for visualization
  - Partial Data: visualize the result of up to 20 SNPs
  - Whole Data: visualize the result of whole SNPs
- Y Axis Scale: specify the unit for Y axis of graph
  - -log10 base: show as -log10 (actual value)
  - P-value: show as actual p-value
- Sort by p-value: sort the calculated p-value by increasing order
  - Default: display in order of SNP positions in chromosome
  - Sort: sort in order of low p-values
- Chart: specify estimated value under the p-value graph
  - OR & CI: odds ratio and its 95% confidence interval
  - AR%: attributable risk %
  - PAR%: population attributable Risk %
- Chart Scale: specify Y axis scale
  - Default: show as analysis result
  - 3.0: set the maximum value at 3.0
- Significant SNP List: SNP list below significance level
  - No: serial number
  - SNP: SNP ID
  - -log10(p): -log10 (p-value)

- Click [Export Sig. SNP List] to show the list of significant SNPs as in <Figure 4-4>.



<Figure 4-4> Statistically significant SNP list

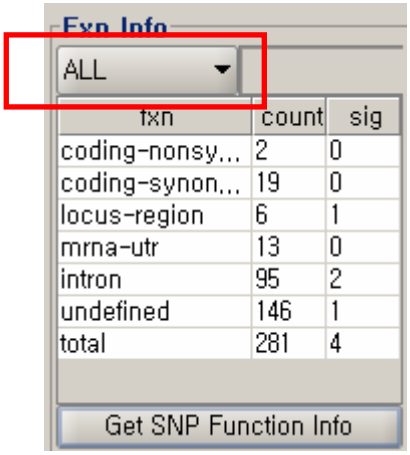
- Click [Export Image] and a window shows as in <Figure 4-5>. Click [OK] after selecting figure file format to save and the saved file is added in "Report" tab of project tree.



<Figure 4-5> Save figure file

- SNP Function Class: specify function class of SNPs for visualization
  - Click [Get SNP Functional Info] to show the function class of SNPs as in <Figure 4-6>. The contents displayed in <Figure 4-6> are the following:

- Function: Defined in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
  - Coding-nonsynonymous
  - Coding-synonymous
  - Intron
  - Mrna-utr
  - Locus-region
  - Undefined: Without locus information
- Total: Number of remaining SNPs after preprocess
- Significant: Number of statistically significant SNPs
- You can display SNPs by function class as in <Figure 4-6>.

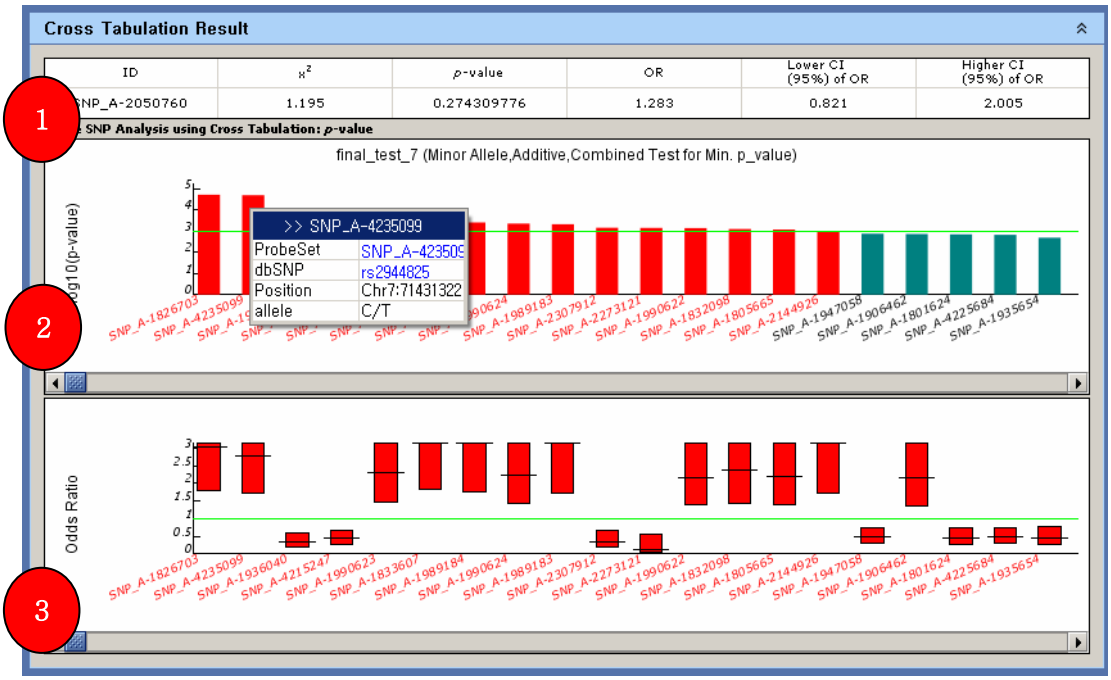


<Figure 4-6> SNP function class information

- Click [Statistics] to show statistical table for the result as shown in <Figure 4-2>.
- Click [Export Annotation] to extract biological annotation information about the significant SNPs. For the details about biological annotation information, please refer to **Chapter 5 Export**..

4.1.2. Cross Tabulation Analysis Control and Result Graph Panel

The graph in the top of <Figure 4-7> shows the p-value of the chi-square test for SNPs. The horizontal line in green indicates the significance level. The graph in the bottom shows Odds Ratio, 95% confidence interval of Odds Ratio, Attributable Risk %, or Population Attributable Risk %. Click a specific SNP to display the analysis result of the SNP. Right-click a specific SNP to display the basic information of the SNP along with dbSNP site connection.

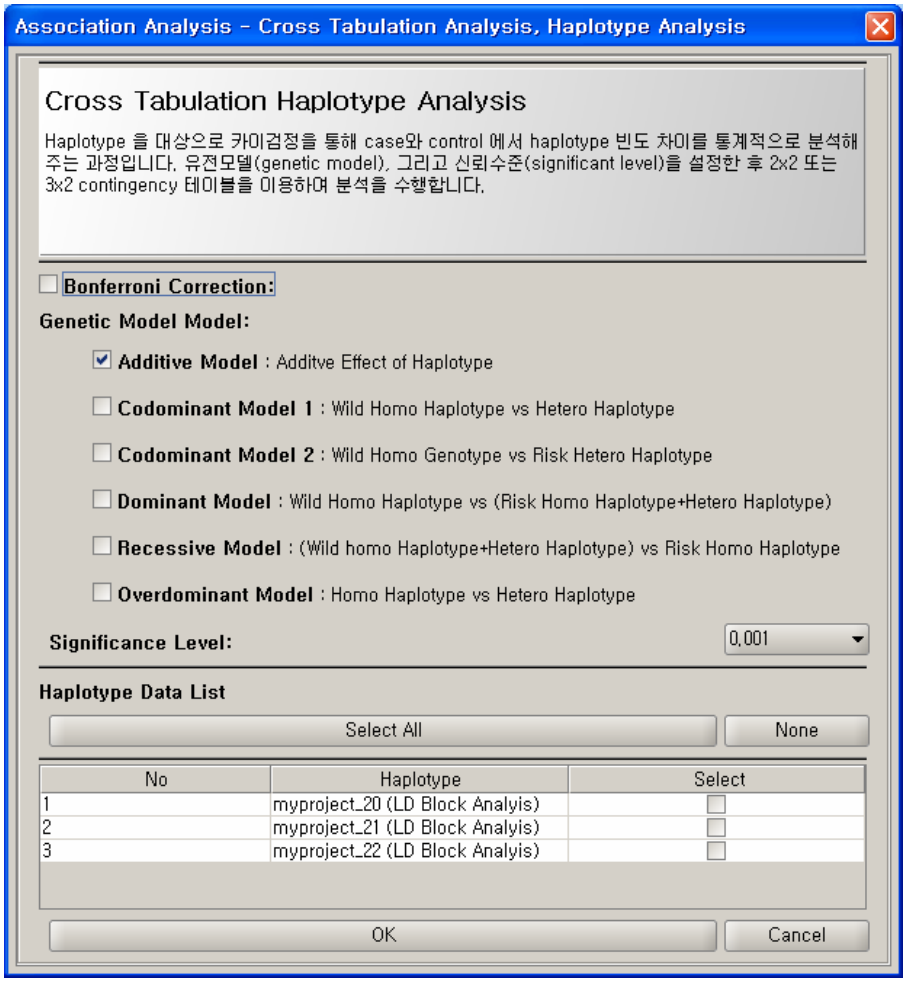


<Figure 4-7> Analysis result graph

4.2. Cross Tabulation Analysis using Haplotype

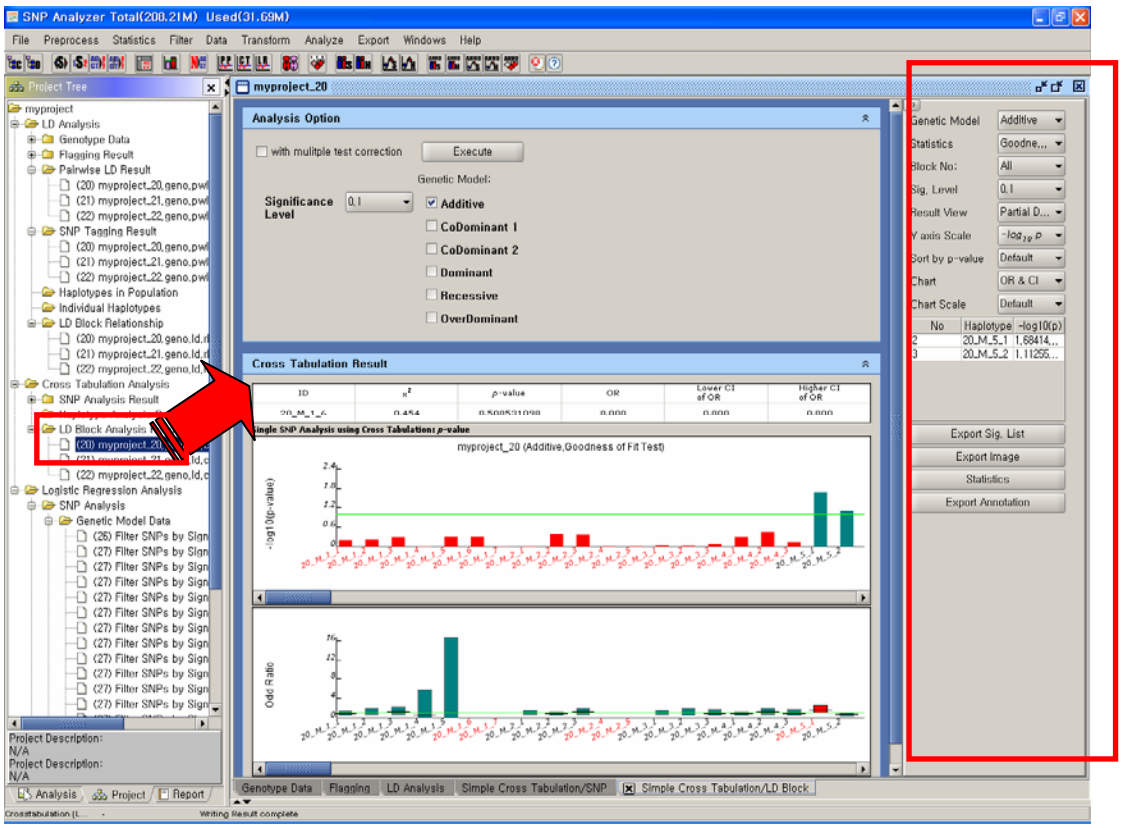
Click [Analyze] > [Association Analysis] > [Cross Tabulation Analysis] > [Haplotype Analysis] to show a window where you can perform case-control analysis using haplotype as in <Figure 4-8>. LD blocking analysis or haplotype estimation is required for the analysis. All the reconstructed haplotypes are automatically analyzed and you can select multi analysis models. For more details about the analysis, please refer to **Appendix-B**.

- Genetic Model
  - Additive Model
  - Codominant Model 1
  - Codominant Model 2
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Estimated Value
  - Odds Ratio
  - Attributable Risk %
  - Population Attributable Risk %



<Figure 4-8> Cross Tabulation Analysis setting window

Set significance level after setting genetic model and select files to analyze. Once analysis is completed, result data is added in project tree. Select one of the analysis results added in project tree to display in graph.



<Figure 4-9> Cross Tabulation Analysis result

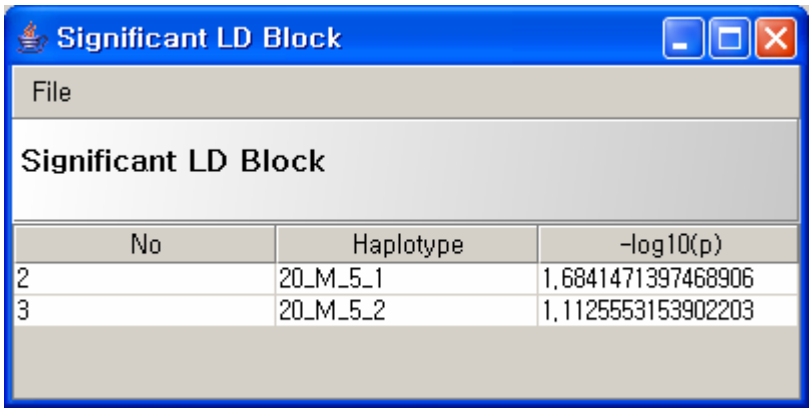
4.2.1. Graph Visualization Control and Result Saving Panel

The right panel in <Figure 4-9> shows the way to visualize analysis result and the list of statistically significant haplotype. The details are the following:

- Genetic Model: specify model used in Analysis
  - Additive Model
  - Codominant Model 1
  - Codominant Model 2
  - Dominant Model
  - Recessive Model
  - Overdominant Model
- Block No: specify LD block number to visualize
- Sig. Level: specify significance level
  - Setting values: 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00001, 0.000001
- X axis Scale: Set number of SNPs for X axis of graph
  - Partial Data: visualize the result of up to 20 SNPs

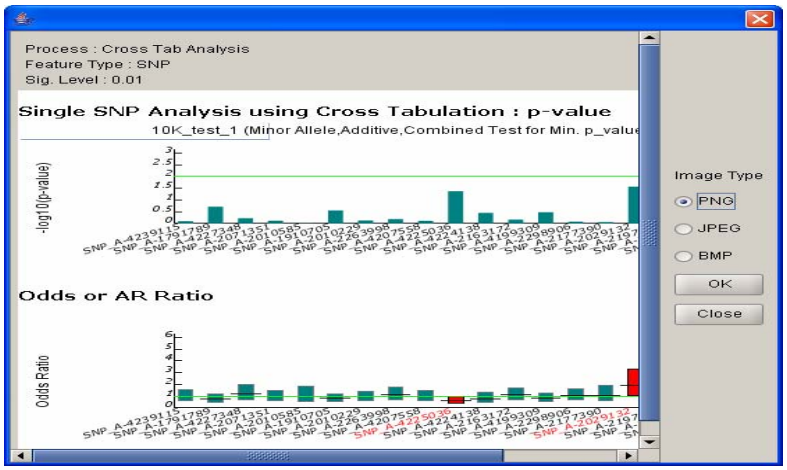
- Whole Data: visualize the result of whole SNPs

- Y Axis Scale: set the unit for Y axis of graph
  - -log10 base: show as -log10 (actual value)
  - p-value: show as actual p-value
- Sort by p-value: sort the calculated p-value by increasing order
  - Default: display in order of SNP position in chromosome
  - Sort: sort in order of low p-value
- Chart: specify estimated value in the bottom of p-value graph
  - OR & CI: odds ratio and its 95% confidence interval
  - AR%: attributable risk %
  - PAR%: population attributable risk %
- Chart Scale: specify Y axis scale
  - Default: show as analysis result
  - 3.0: set the maximum value at 3.0
- Significant Haplotype List: list of haplotype below significance level
  - No: serial number
  - Haplotype: haplotype ID
  - -log10(p): -log10(p-value)
- Click [Export Sig. Haplotype List] to show the list of significant haplotypes as a pop-up window as in <Figure 4-10>.



<Figure 4-10> List of haplotype extracted statistically significant

- Click [Export Image] and a window shows as in <Figure 4-11>. Click [OK] after selecting a figure file format to save and the saved file in the “Report” tab of project tree.



<Figure 4-11> Save figure file

- Click [Export Annotation] to extract biological annotation information of the significant haplotypes. For the details about biological annotation information extraction, please refer to **Chapter 5 Export..**

4.2.2. Cross Tabulation Analysis Control and Result Graph Panel

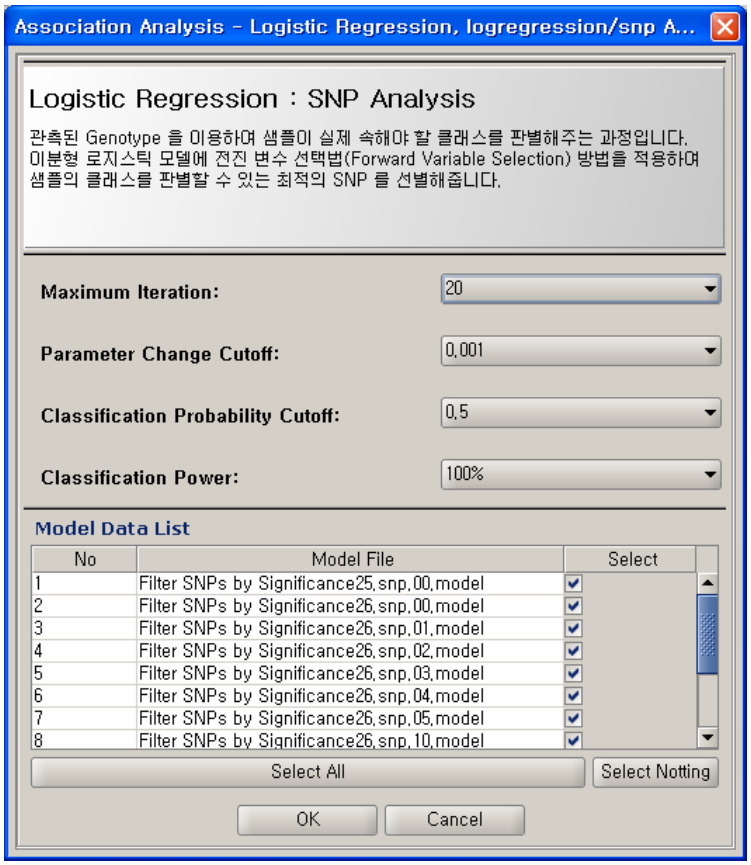
The graph in <Figure 4-12> shows p-values of chi-square test. The horizontal line in green indicates the significance level. The graph below the p-value shows Odds Ratio, Attributable Risk %, or Population Attributable Risk %. Odds Ratio and its 95% confidence intervals are displayed simultaneously. Click a specific haplotype to display the analysis result in table.

<Figure 4-12> Save figure file

4.3. Logistic Regression Analysis Using SNP

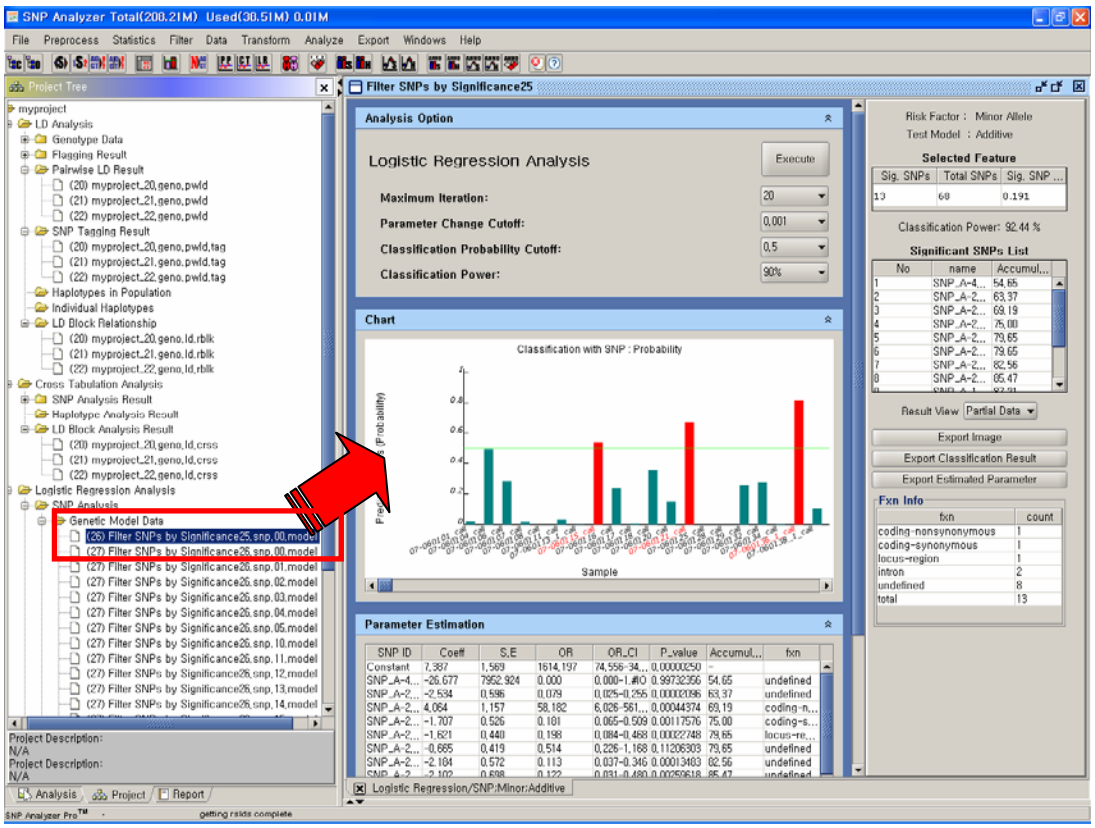
Click [Analyze] > [Association Analysis] > [Logistic Regression] > [SNP Analysis] to show a

window where you can set analysis parameters as in <Figure 4-13>. (For the details about parameters used for analysis, please refer to **Appendix-B**). A file with *.model extension is needed to perform logistic regression analysis. For the details about model file creation, please refer to **Chapter 6 Transformation.** Click [OK] after selecting a model file to analyze. Once analysis is completed, the result data is automatically added in project tree.



<Figure 4-13> Logistic Regression Analysis setting window

Select and double-click one of the analysis result files (*.plog or *.ilog) in project tree to show in graph as in <Figure 4-14>.

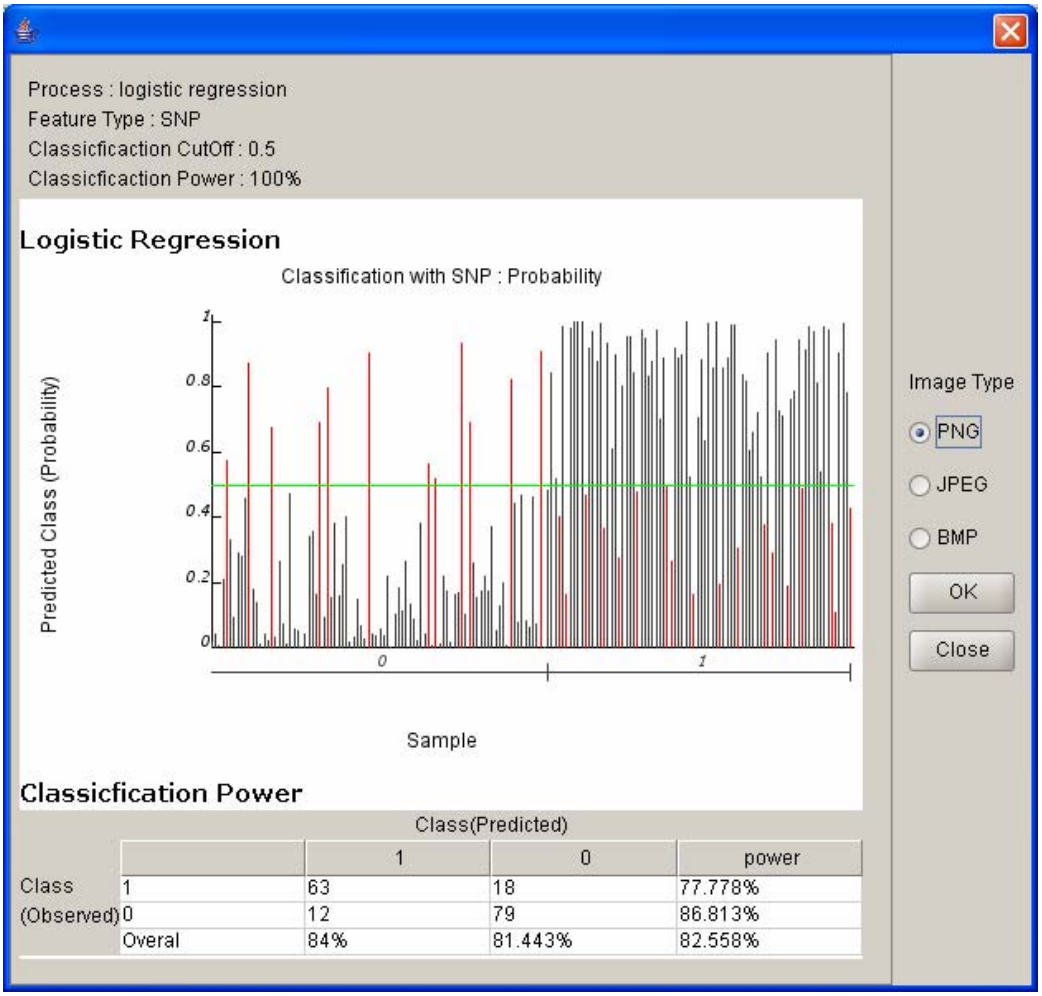


<Figure 4-14> Logistic Regression Analysis result

4.3.1. Graph Visualization Control and Result Saving Panel

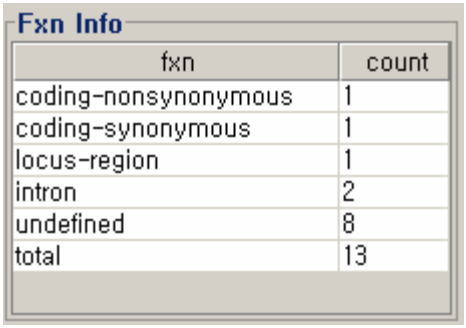
The right panel in <Figure 4-14> is where you can control the visualization parameters. The details are the following:

- Risk Factor: allele specificed as risk factor
- Test Model: genetic model to analyze in model file
- Selected Feature
  - Selected: number of SNPs selected as classification marker
  - Total: total number of SNPs used in analysis
- Classification Power: total classification power of SNPs selected as classification marker
- Accumulated Power: accumulated classification power of SNPs selected as classification marker
- X axis Scale: specify number of samples shown in X axis of graph
  - Partial Data: visualize the result of up to 20 samples
  - Whole Data: visualize the result of entire sample
- Click [Export Image] to save the result graph as figure file as in <Figure 4-15>. Saved figure files are added in project tree.



<Figure 4-15> Sample determining result and save in figure File

- SNP Function Class: function class information of SNPs
  - Click [Get SNP Function Info] to display the function class of SNPs as in <Figure 4-16>. The contents displayed in <Figure 4-16> are the following:
  - Function: Defined in dbSNP(http://www.ncbi.nlm.nih.gov/projects/SNP/)
    - Coding-nonsynonymous
    - Coding-synonymous
    - Intron
    - Mrna-utr
    - Locus-region
    - Undefined: without locus information
    - Total: number of total function classes



Fxn Info

fxn	count
coding-nonsynonymous	1
coding-synonymous	1
locus-region	1
intron	2
undefined	8
total	13

<Figure 4-16> SNP function class information

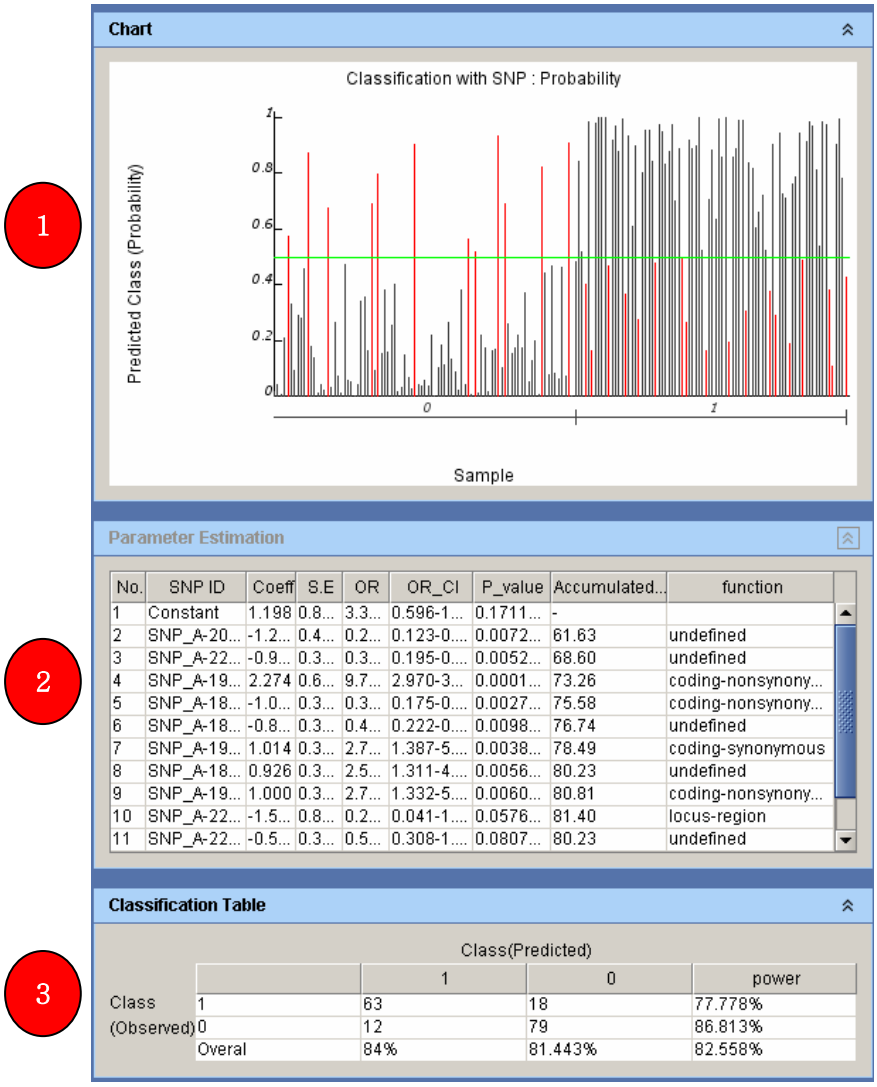
- Click [Export Classification Result] to view the classification table. For the details about classification result, please refer to **Chapter 5 Export.**
- Click [Export Parameter Estimate] to view the values of the coefficients that are estimated by logistic regression analysis. For the details about analysis result, please refer to **Chapter 5 Export.**

4.3.2. Logistic Regression Analysis Control and Result Graph Panel

The upper part of the graph in <Figure 4-17> shows the each sample's classification result. The threshold probability (default=0.5) is shown in green line. Correctly classified sample is shown in green and incorrectly classified sample is shown in red. The estimated values of the selected markers' coefficients are shown in the "Parameter Estimation" table. Descriptions of each item are the following:

- SNP ID
- Coeff: estimated value of the coefficient that corresponds to each SNP in logistic regression
- S.E: standard error of estimated coefficient
- OR: adjusted Odds Ratio
- OR_CI: 95% confidence interval of OR
- P_value: p_value of the estimated coefficient
- Accumulated Power: accumulated classification power
- Function: function class of the SNP

The total classification result for sample is shown in "Classification Table". It shows the whole classification power with the correctly and incorrectly classified number of samples.



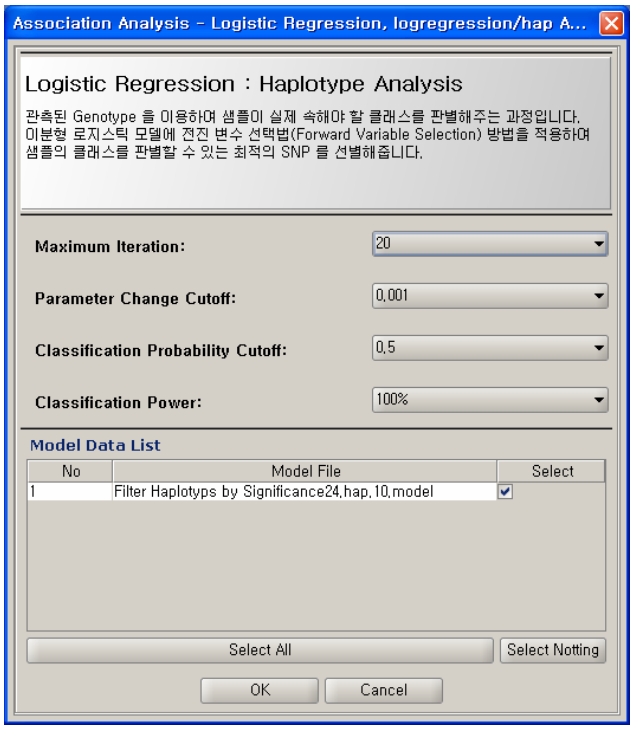
No.	SNP ID	Coeff	S.E	OR	OR_CI	P_value	Accumulated...	function
1	Constant	1.198	0.8...	3.3...	0.596-1...	0.1711...	-	
2	SNP_A-20...	-1.2...	0.4...	0.2...	0.123-0....	0.0072...	61.63	undefined
3	SNP_A-22...	-0.9...	0.3...	0.3...	0.195-0....	0.0052...	68.60	undefined
4	SNP_A-19...	2.274	0.6...	9.7...	2.970-3...	0.0001...	73.26	coding-nonsynony...
5	SNP_A-18...	-1.0...	0.3...	0.3...	0.175-0....	0.0027...	75.58	coding-nonsynony...
6	SNP_A-18...	-0.8...	0.3...	0.4...	0.222-0....	0.0098...	76.74	undefined
7	SNP_A-19...	1.014	0.3...	2.7...	1.387-5....	0.0038...	78.49	coding-synonymous
8	SNP_A-18...	0.926	0.3...	2.5...	1.311-4....	0.0056...	80.23	undefined
9	SNP_A-19...	1.000	0.3...	2.7...	1.332-5....	0.0060...	80.81	coding-nonsynony...
10	SNP_A-22...	-1.5...	0.8...	0.2...	0.041-1....	0.0576...	81.40	locus-region
11	SNP_A-22...	-0.5...	0.3...	0.5...	0.308-1....	0.0807...	80.23	undefined

		1	0	power
Class (Observed)	1	63	18	77.778%
	0	12	79	86.813%
	Overal	84%	81.443%	82.558%

<Figure 4-17> Analysis result graph

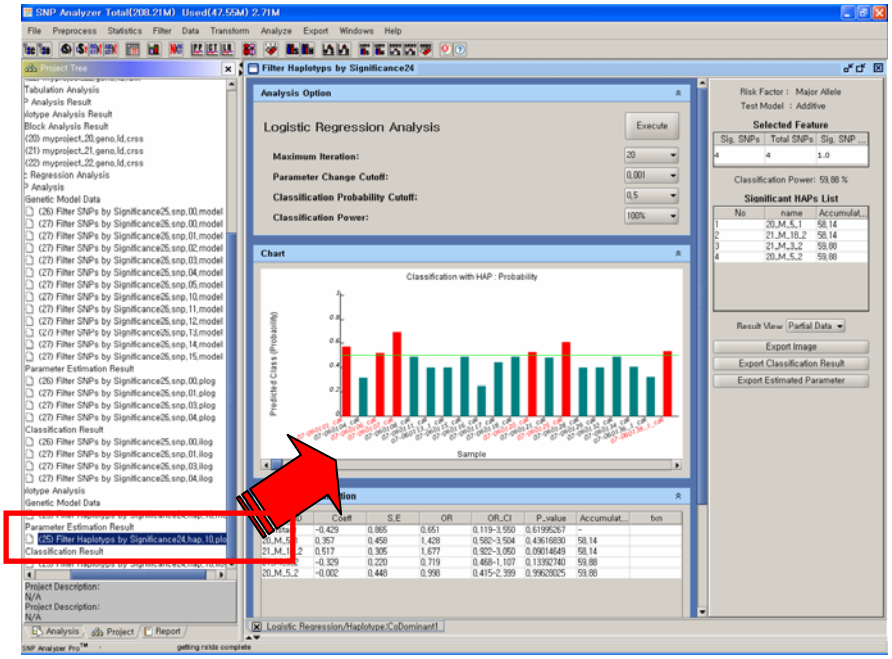
4.4. Logistic Regression Analysis using haplotype

Click [Analyze] > [Association Analysis] > [Logistic Regression] > [Haplotype Analysis] to show a window where you can set analysis parameters as in <Figure 4-18>. (For the details about parameters used for analysis, please refer to **Appendix-B**). A file with *.model extension is needed to perform logistic regression analysis. For the details about model file creation, please refer to **Chapter 6 Transformation.** Click [OK] after selecting a model file to analyze. Once analysis is completed, the result data is automatically added in project tree.



<Figure 4-18> Logistic Regression Analysis setting window

Select and double-click one of the analysis result files (*.plog or *.ilog) added in project tree to display the analysis result in graph as in <Figure 4-19>.



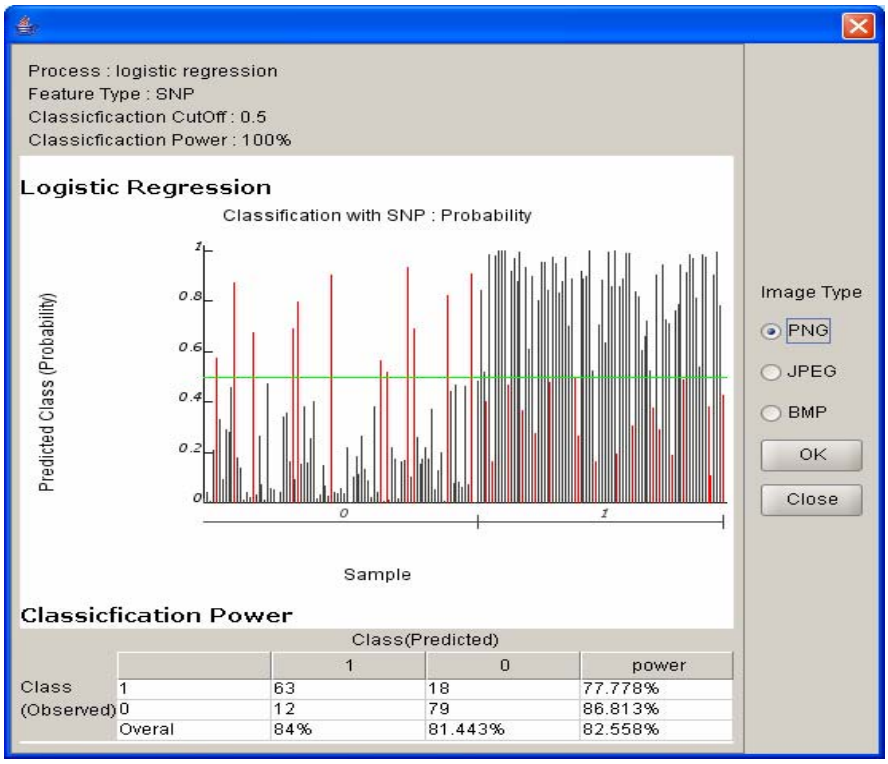
<Figure 4-19> Logistic Regression Analysis result

4.4.1. Graph Visualization Control and Result Saving Panel

The right panel in <Figure 4-19> is where you can control the visualization parameters. The details are the following:

- Risk Factor: significant haplotype extracted from the cross tabulation analysis
- Test Model: genetic model to analyze in model file
- Selected Feature
  - Selected: number of haplotypes selected as classification marker
  - Total: total number of haplotypes used in analysis
- Classification Power: total classification power of haplotypes selected as classification marker
- Accumulated Power: accumulated classification power of haplotypes selected as classification marker
- X axis Scale: specify number of samples shown in X axis of graph
  - Partial Data: visualize the result of up to 20 samples
  - Whole Data: visualize the result of entire sample

Click [Export Image] to save the result graph as figure file as in <Figure 4-20>. Saved files are added in project tree.



<Figure 4-20> Save sample classification result in figure file

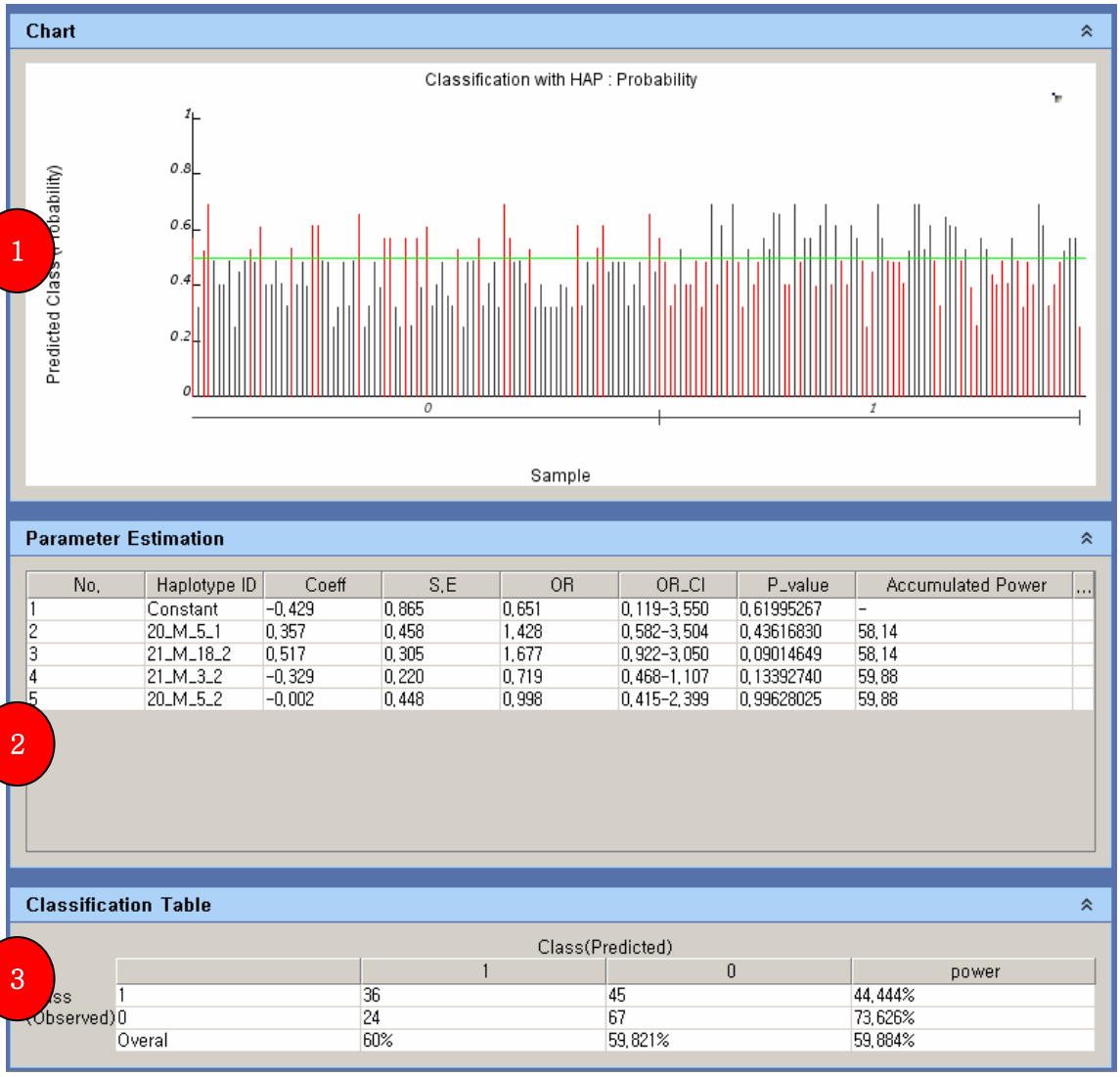
- Click [Export Classification Result] to view the whole classification result. For the details about classification result, please refer to **Chapter 5 Export.**
- Click [Export Parameter Estimate] to view the analysis results for each classification marker. For the details about analysis result, please refer to **Chapter 5 Export.**

4.4.2. Logistic Regression Analysis Control and Result Graph Panel

The upper part of the graph in <Figure 4-21> shows the each sample's classification result. The threshold probability (default=0.5) is shown in green line. Correctly classified sample is shown in green and incorrectly classified sample is shown in red. The estimated values of the selected markers' coefficients are shown in the "Parameter Estimation" table. Descriptions of each item are the following:

- Haplotype ID
- Coeff: estimated value of the coefficient that corresponds to each haplotype in logistic regression
- S.E: standard error of estimated coefficient
- OR: adjusted Odds Ratio
- OR_CI: 95% confidence interval of OR
- P_value: p_value of the estimated coefficient
- Accumulated Power: accumulated classification power

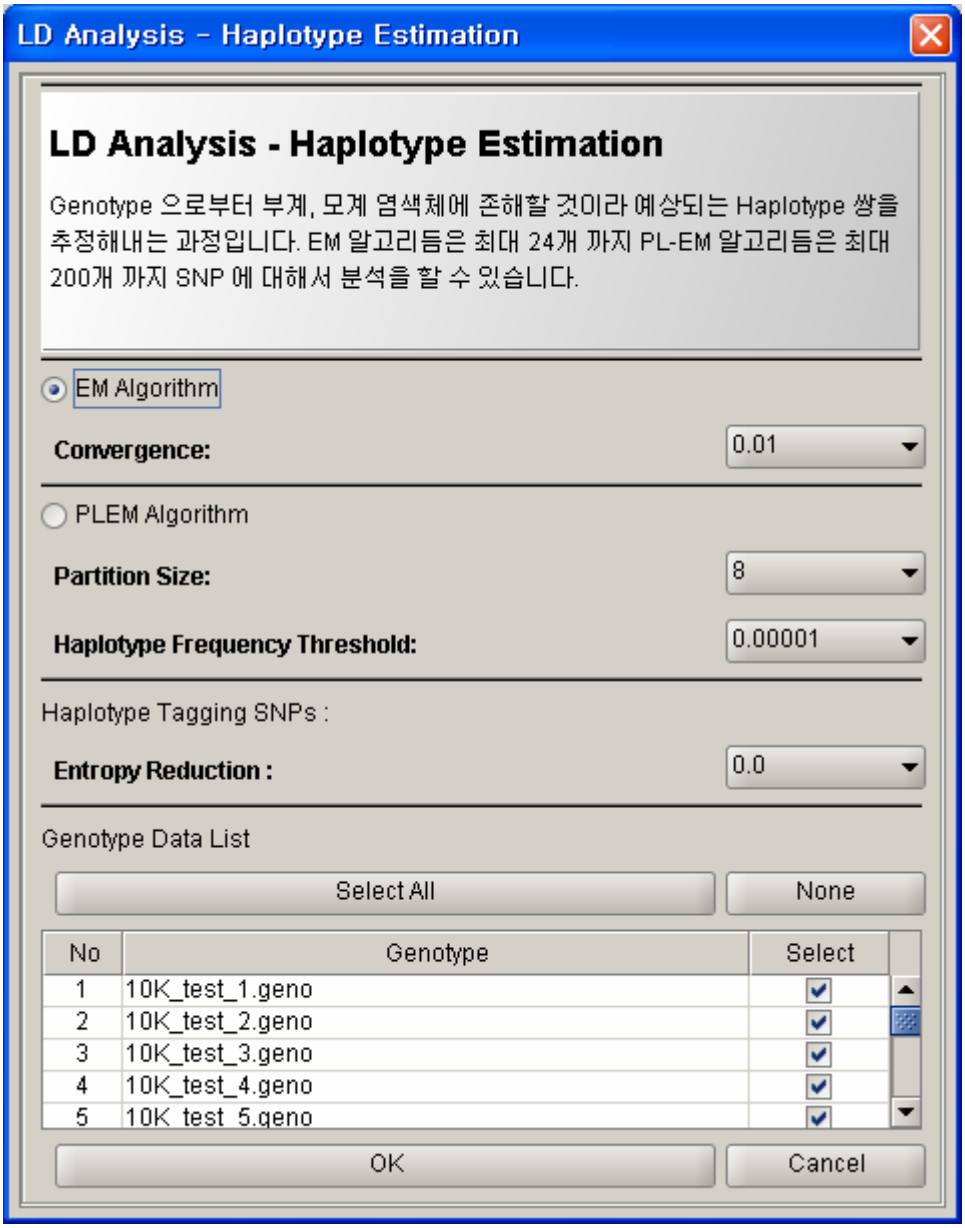
The total classification result for sample is shown in "Classification Table". It shows the whole classification power with the correctly and incorrectly classified number of samples.



<Figure 4-21> Analysis result graph

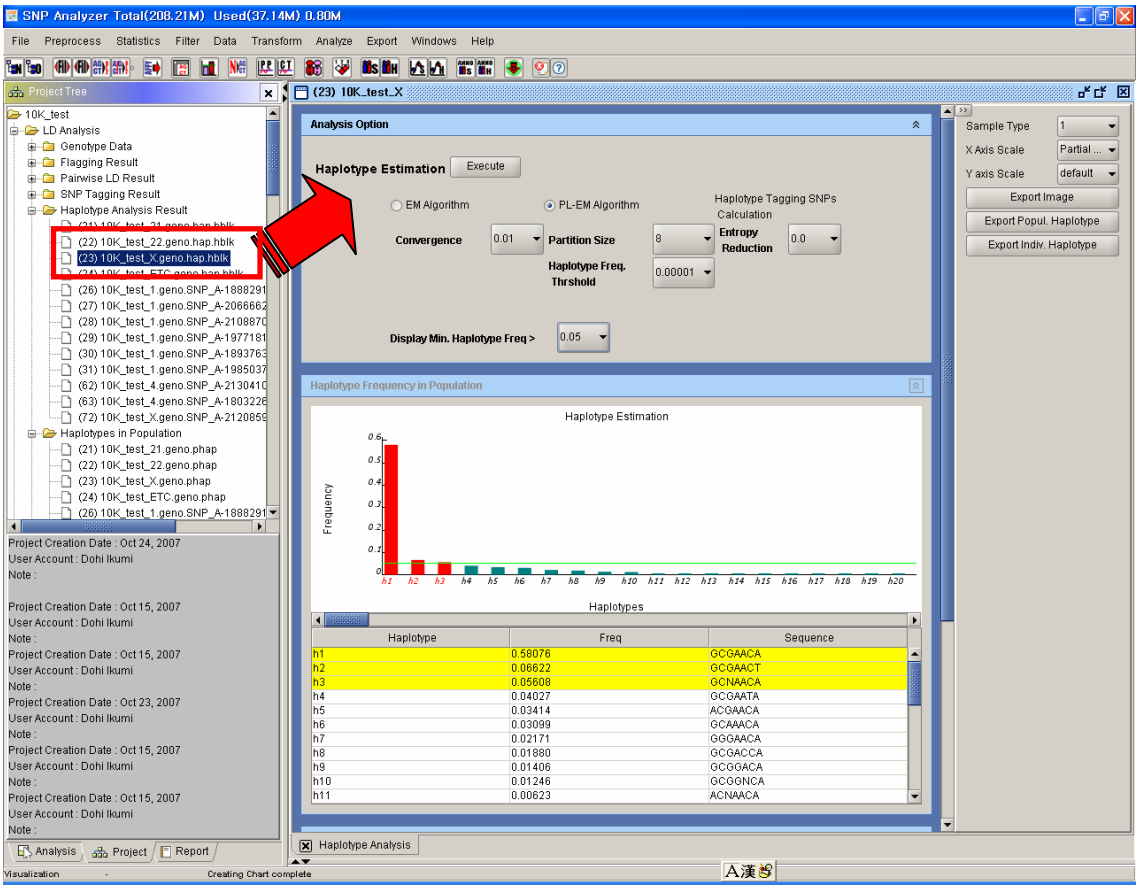
4.5. Haplotype Estimation

You can reconstruct haplotypes using the genotype data. EM algorithm and PL-EM algorithms are used for the haplotype reconstruction. Click [Analyze] > [LD Analysis] > [Haplotype Estimation] to show the window where you can set parameters required to perform algorithm as shown in <Figure 4-22>. Click [OK] after selecting a genotype to analyze in “Genotype Data List”. When the analysis is completed, result data is automatically added in project tree.



<Figure 4-22> Set haplotype estimation parameters

Select and double-click one of the analysis results in project tree to show the analysis result in graph and table as in <Figure 4-23>.



<Figure 4-23> Haplotype Estimation Analysis result

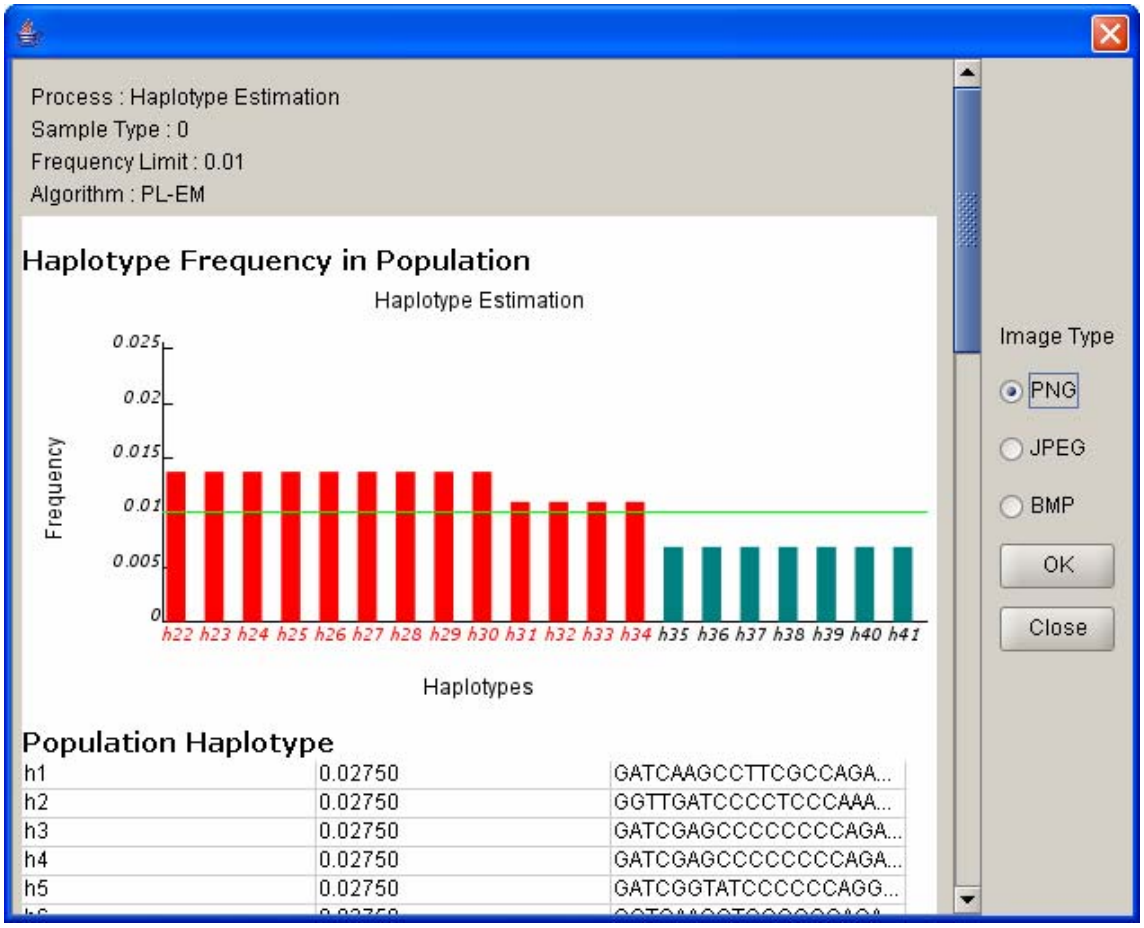
4.5.1. Graph Visualization Control and Result Saving Panel

The right panel in <Figure 4-23> shows the visualization control. The details are the following:

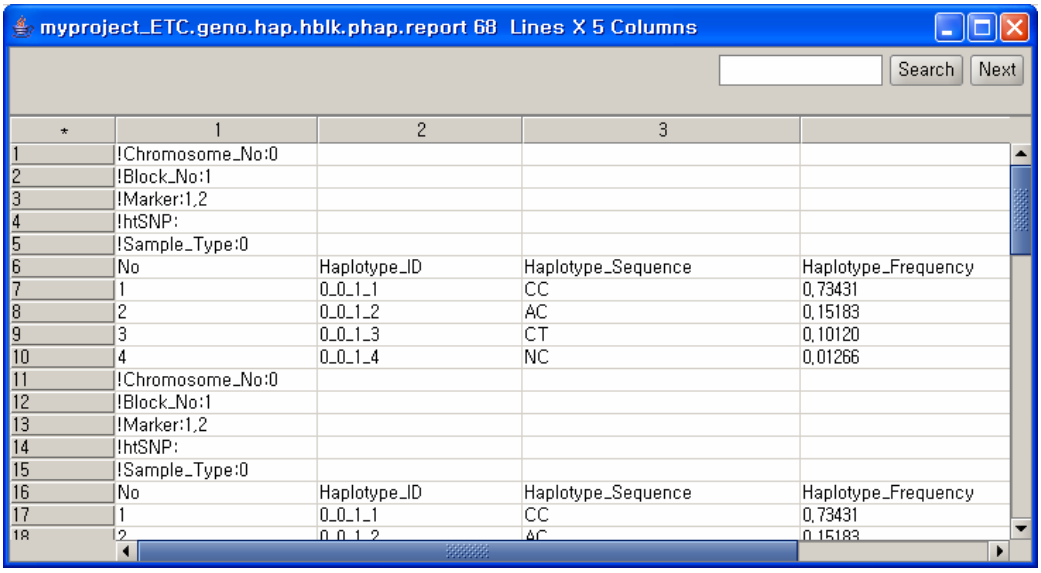
- Sample Type: select input sample
  - 0: control sample
  - 1: case sample
  - M: total of control sample and case sample
- X axis Scale: specify number of haplotypes to show in X axis of graph
  - Partial Data: visualize up to 20 haplotypes
  - Whole Data: visualize all the reconstructed haplotypes
- Y Axis Scale: set the unit for Y axis of graph
  - Default: set the maximum value of Y axis to the largest haplotype frequency
  - Max 0.5: set the maximum value of Y axis to 0.5
  - Max 1.0: set the maximum value of Y axis to 1.0
- Click [Export Image] and a window as in <Figure 4-24> is displayed. The saved figure

files are automatically added in “Report” tab of project tree.

- Click [Export Popul. Haplotype] to show the haplotypes reconstructed in the specified sample and haplotype frequencies in table as in <Figure 4-25>.
- Click [Export Indiv. Haplotype] to show the estimated haplotype set of each individual and estimation accuracy in table as in <Figure 4-26>.

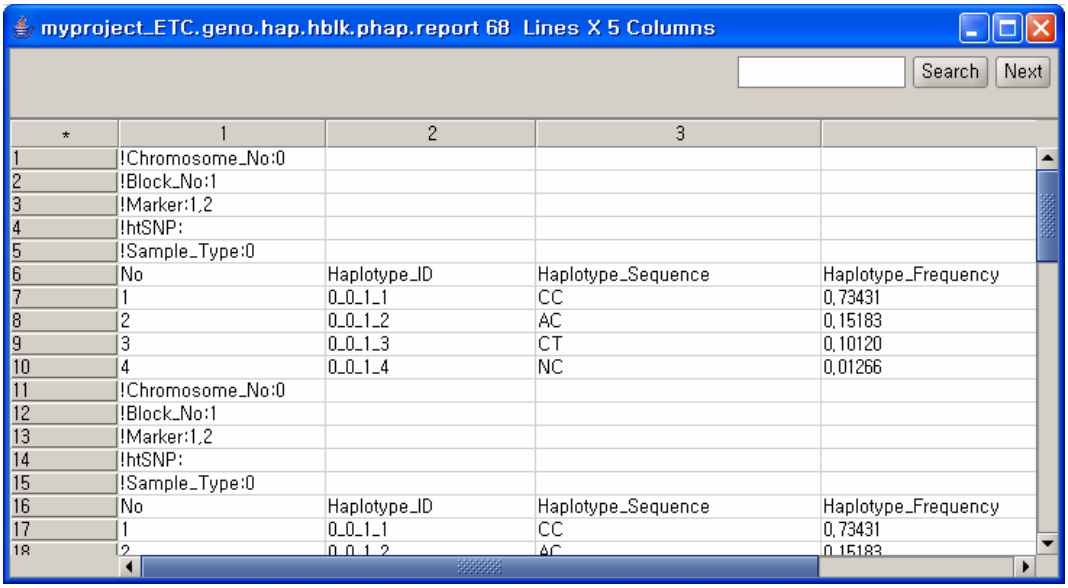


<Figure 4-24> Save haplotype estimation result



*	1	2	3	
1	!Chromosome_No:0			
2	!Block_No:1			
3	!Marker:1,2			
4	!htSNP:			
5	!Sample_Type:0			
6	No	Haplotype_ID	Haplotype_Sequence	Haplotype_Frequency
7	1	0_0_1_1	CC	0,73431
8	2	0_0_1_2	AC	0,15183
9	3	0_0_1_3	CT	0,10120
10	4	0_0_1_4	NC	0,01266
11	!Chromosome_No:0			
12	!Block_No:1			
13	!Marker:1,2			
14	!htSNP:			
15	!Sample_Type:0			
16	No	Haplotype_ID	Haplotype_Sequence	Haplotype_Frequency
17	1	0_0_1_1	CC	0,73431
18	2	0_0_1_2	AC	0,15183

<Figure 4-25> Estimated haplotype of sample



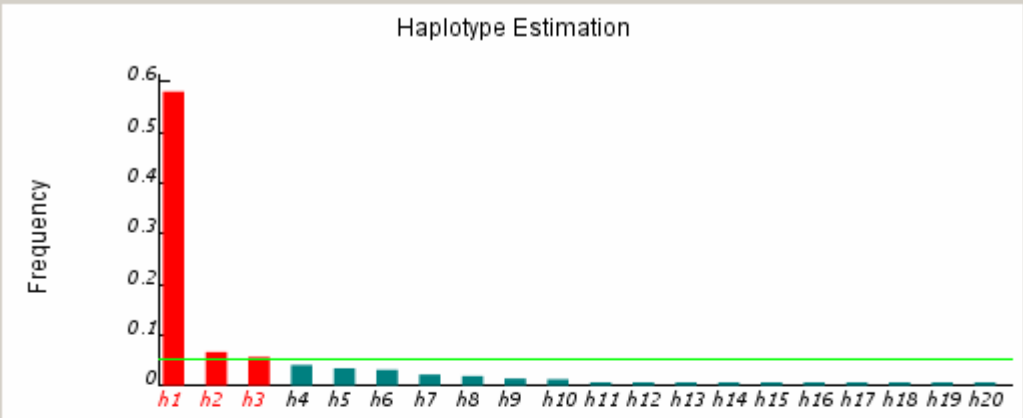
*	1	2	3	
1	!Chromosome_No:0			
2	!Block_No:1			
3	!Marker:1,2			
4	!htSNP:			
5	!Sample_Type:0			
6	No	Haplotype_ID	Haplotype_Sequence	Haplotype_Frequency
7	1	0_0_1_1	CC	0,73431
8	2	0_0_1_2	AC	0,15183
9	3	0_0_1_3	CT	0,10120
10	4	0_0_1_4	NC	0,01266
11	!Chromosome_No:0			
12	!Block_No:1			
13	!Marker:1,2			
14	!htSNP:			
15	!Sample_Type:0			
16	No	Haplotype_ID	Haplotype_Sequence	Haplotype_Frequency
17	1	0_0_1_1	CC	0,73431
18	2	0_0_1_2	AC	0,15183

<Figure 4-26> Estimated individual haplotype

4.5.2. Haplotype Estimation Control and Result Graph Panel

The upper graph in <Figure 4-27> shows the haplotype frequencies estimated in the sample. The corresponding haplotypes and frequencies are shown in middle. The table on the bottom shows the estimated haplotype of each individual, estimation accuracy and sample type.

Haplotype Frequency in Population



Haplotype	Freq	Sequence
h1	0.58076	GCGAACA
h2	0.06622	GCGAACT
h3	0.05608	GCNAACA
h4	0.04027	GCGAATA
h5	0.03414	ACGAACA
h6	0.03099	GCAAACA
h7	0.02171	GGGAACA
h8	0.01880	GCGACCA
h9	0.01406	GCGGACA
h10	0.01246	GCGGNCA
h11	0.00623	ACNAACA

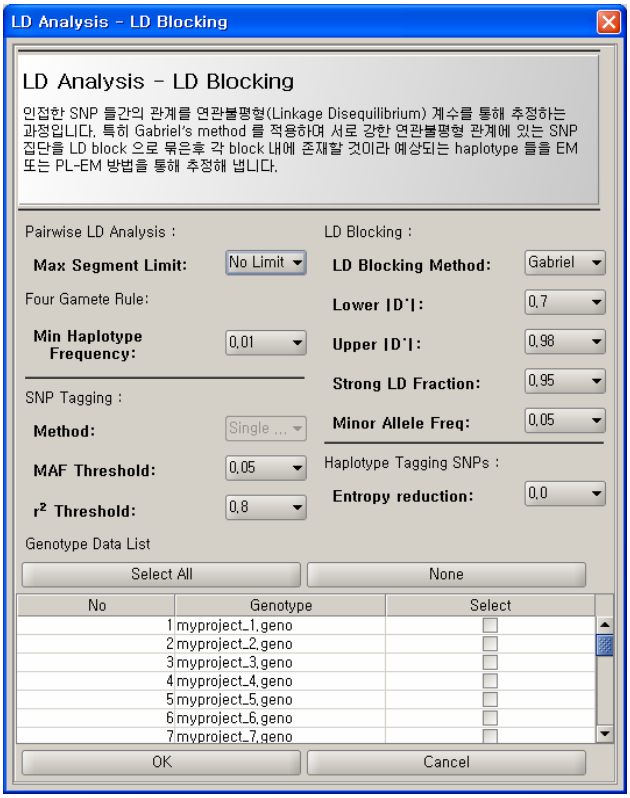
Individual Haplotypes

Individual ID	Haplotype1	Haplotype1	Accuracy	Sample Type
01-051008_call	h5	h1	1.00000	1
01-051102_call	h1	h8	1.00000	1
01-051111_call	h6	h1	1.00000	1
01-051206_call	h1	h1	1.00000	1
01-051210_call	h1	h1	1.00000	1
01-060108_call	h1	h1	1.00000	1
01-060118_call	h1	h4	1.00000	1
01-060204_call	h1	h1	1.00000	1
01-060206_1_c...	h5	h5	1.00000	1
01-060407_1_c...	h1	h1	1.00000	1
01-060414_call	h1	h19	0.99998	1
01-060417_call	h1	h1	1.00000	1
01-060419_call	h1	h1	1.00000	1
01-060509_call	h1	h1	1.00000	1

<Figure 4-27> Estimated haplotype result graph

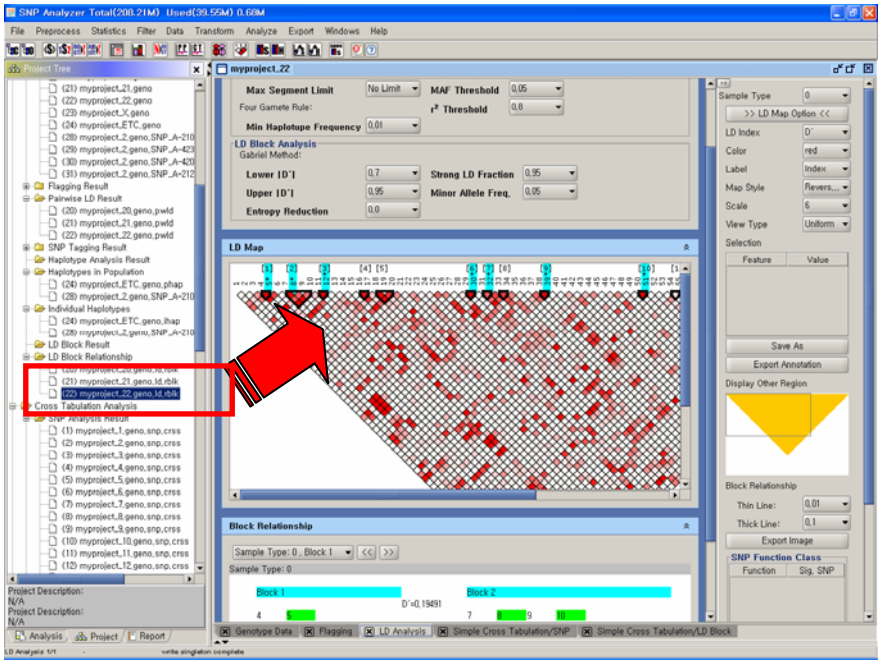
4.6. LD Blocking with Gabriel's Method

SNPs that are in strong linkage disequilibrium can be grouped into one block. Click [Analyze] > [LD Analysis] > [LD Blocking with Gabriel's Method] to show the window where you can perform LD block analysis as in <Figure 4-28>. Click [OK] after selecting a genotype to analyze from "Genotype Data List". When the analysis is completed, the result data is automatically added in the project tree.



<Figure 4-28> Set LD block analysis parameters

Select and double-click one of the analysis results added in project tree to display the analysis result in graph and table format as in <Figure 4-29>.



<Figure 4-29> LD blocking analysis result

4.6.1. LD Map Visualization Control and Result Saving Panel

The right panel in <Figure 4-29> shows the visualization control and information about SNPs that form LD blocks. The details are the following:

- Sample Type: select sample type
  - 0: control sample
  - 1: case sample
  - M: total of control sample and case sample
- Click [>>LD Map Option<<] and a window where you can control the LD Map visualization appears as in <Figure 4-30>. Descriptions for each item are the following:

  ①: Change the shape of LD Map: “Reverse Triangle” and “Lower Diagonal”

  ②: Distance between SNPs: Physical Distance” and “Uniform Distance”

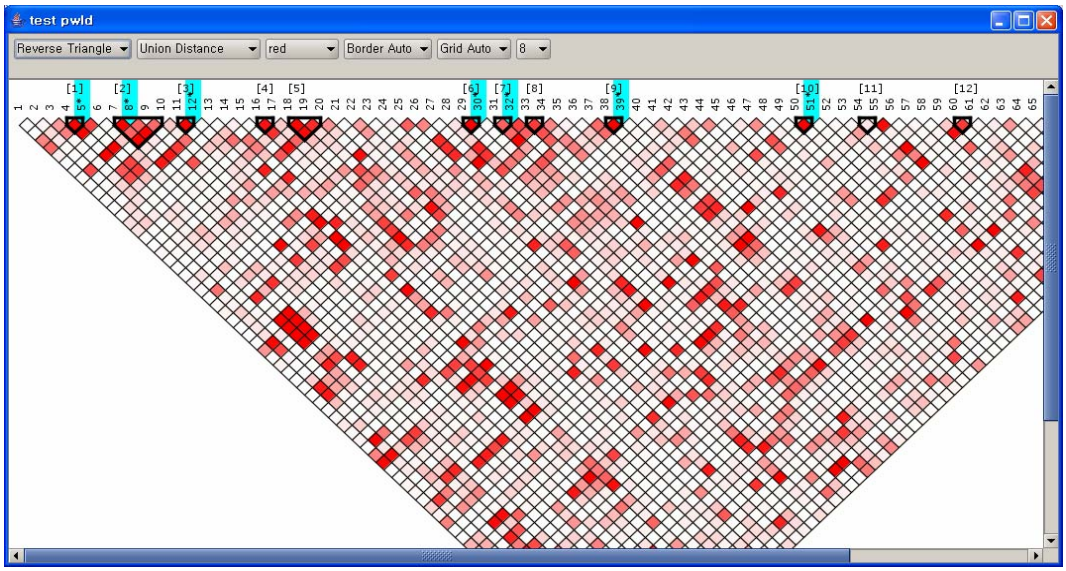
  ③: Change the color of LD: “Red”, “Green”, and “Blue”

  ④: Select “On” to show the boundary line of the square showing the D’ value or select “Off” otherwise.

  ⑤: Select “On” to show the specified area of the SNP pair of which pairwise LD value is not calculated or select “Off” otherwise.

  ⑥: Control the size of LD: “1”, “2”, “4”, “8”, and “16” (each number indicates the number of times the figure size based on “1”.

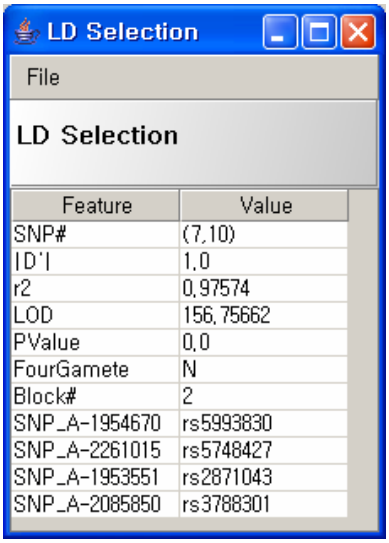
1 2 3 4 5 6



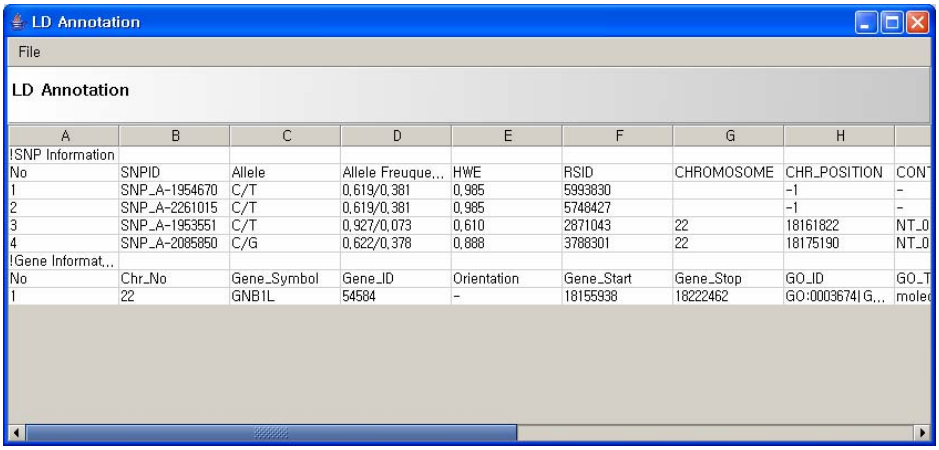
<Figure 4-30> LD map control interface

- LD Index: |D'| or R^2
- Color: change the color of LD Map: "Red", "Green", and "Blue"
- Label: SNP identifier shown above LD Map
  - Index: serial number starting from 1
  - Marker ID: show SNP ID with index
  - None: do not show
- Map Style: change the shape of LD Map: "Reverse Triangle" and "Lower Diagonal"
- Scale: control the size of LD Map: "1", "2", "4", "8", and "16" (each number indicates the number of times the Figure size based on "1".)
- View Type: displayed distance between SNPs: "Physical Distance" and "Uniform Distance"
- SNP Pair & Block Info: SNP and block information.
  - SNP Index: SNP#
  - LD Index Value: |D'|, R^2
  - Chi-squared value for the significance level of D': LOD-Score
  - Independence chi square test result between adjacent SNPs: p-value
  - Four Gamete: Y or No
  - LD block number: Block#
  - SNP ID and dbSNP #rs within the block
- Click [Export SNP Pair & Block Info] to show the window as in <Figure 4-31> where you can save information about the selected SNP pair and block.
- Click [Export Annotation] to view the window that shows the annotation information about

SNPs as in <Figure 4-32>. Saving the annotation information automatically adds the result in the “Report” tab in project tree.

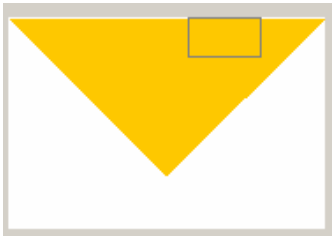


<Figure 4-31> SNP Pair and Block Information

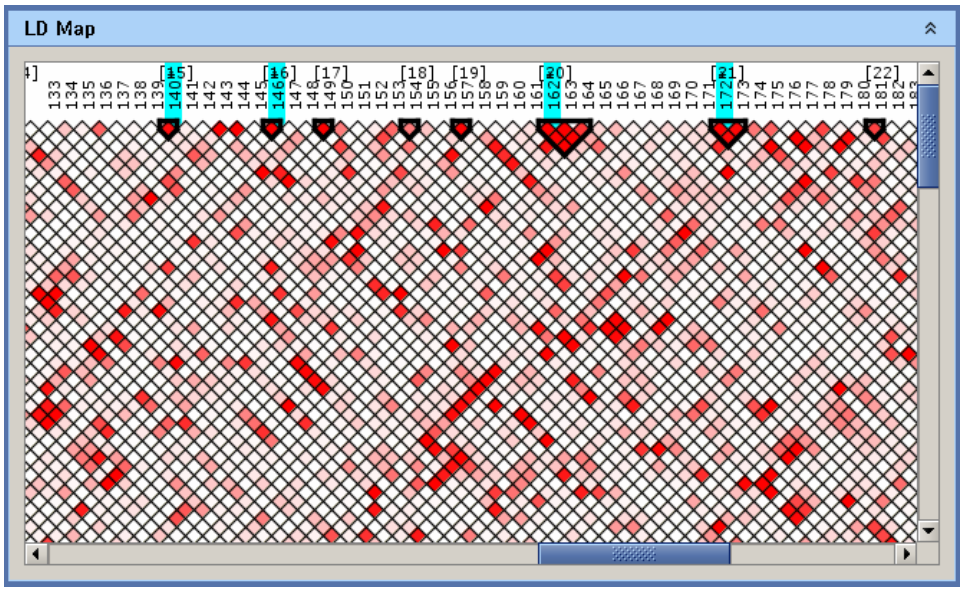


<Figure 4-32> SNP and Chromosome Annotation Information

- Move the square area in gray in “Displaying Region” as in <Figure 4-33> and you can view LD Map of the moved square area of the screen as in <Figure 4-34>.

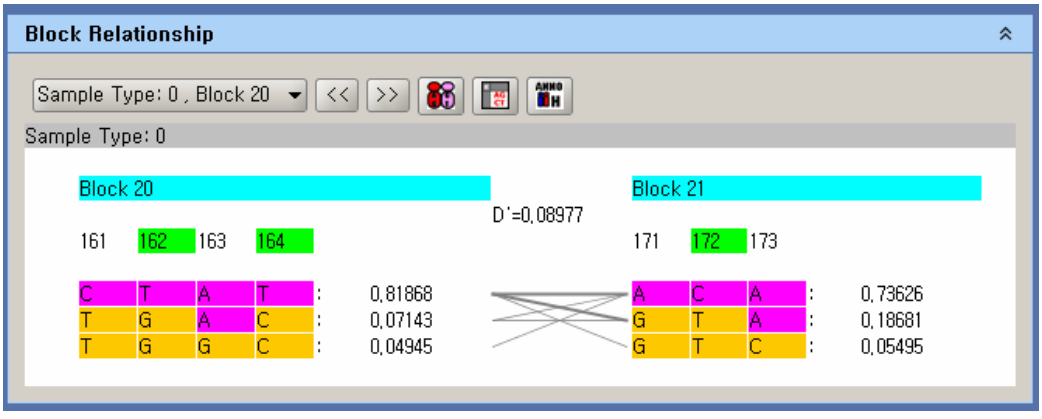


<Figure 4-33> Visualization area move panel



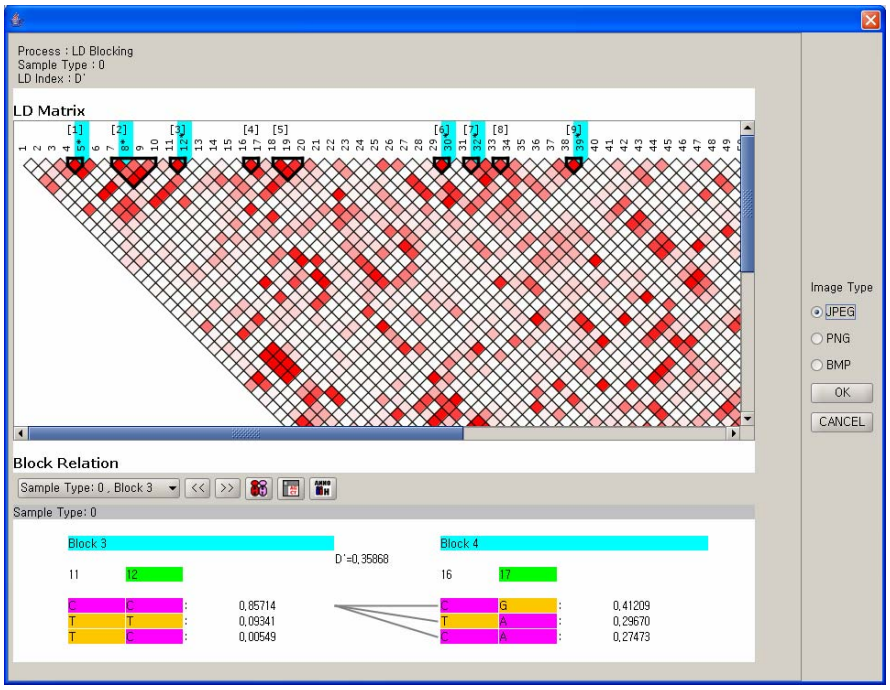
<Figure 4-34> Moved LD map screen

- Block Relationship: set the line thickness according to the crossover rate between estimated haplotypes in two adjacent LD blocks. <Figure 4-35> shows the result according to the thickness of lines set.



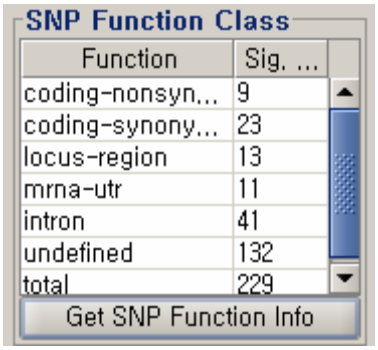
<Figure 4-35> Block relationship

- Click [Export Image] and interface where you can save the LD Map for the selected area in image file appears as in <Figure 4-36>. Click [OK] after selecting a figure file format for saving. The saved image is automatically added in the “Report” tab in project tree.



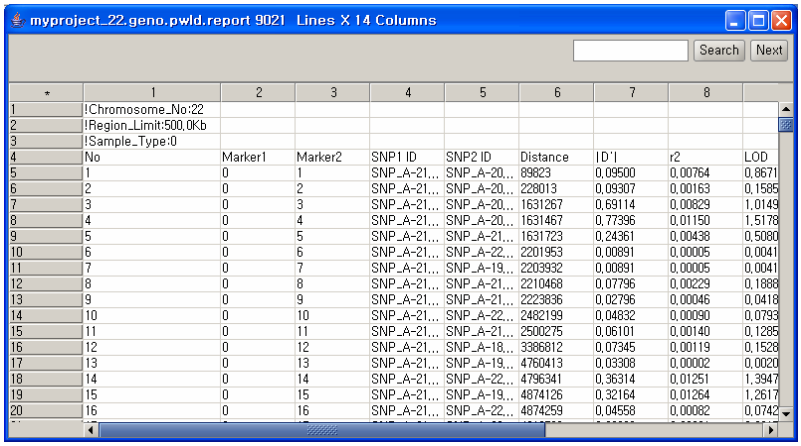
<Figure 4-36> Save LD Map image

- SNP Function Class: click [Get SNP Function Info] to show function class of SNPs as in <Figure <4-37>.
  - Function: defined in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
    - Coding-nonsynonymous
    - Coding-synonymous
    - Intron
    - Mrna-utr
    - Locus-region
    - Undefined: without locus information
    - Total SNP: total number of SNPs in a specified class



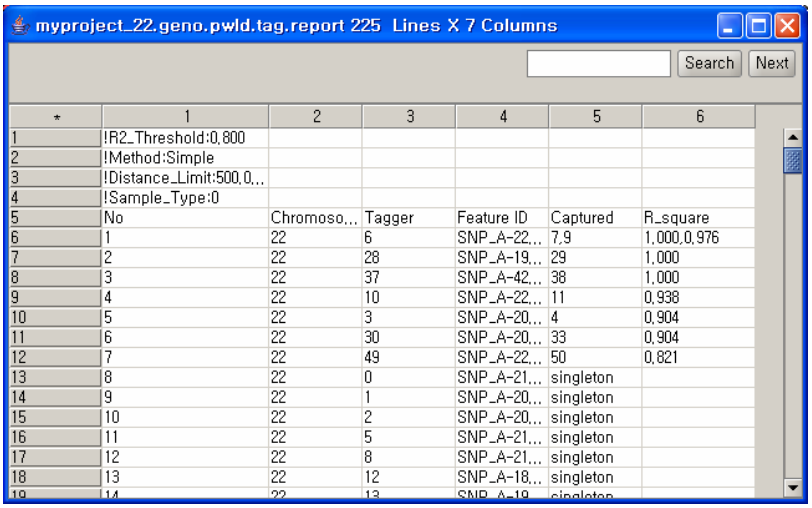
<Figure 4-37> SNP functional class

- Click [Export Pairwise LD] to show the results of LD calculation between pairwise SNPs in table format as in <Figure 4-38>.
- Click [Export Tagging SNPs] to show the tagging SNPs in table format as in <Figure 4-39>.
- Click [Export Block Relationship] to show the haplotype relationships in two adjacent LD blocks in table format as in <Figure 4-40>.



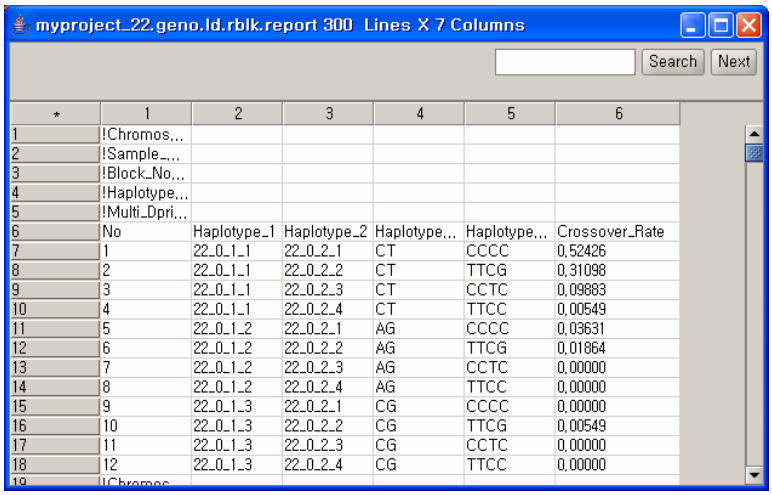
*	1	2	3	4	5	6	7	8	
1	!Chromosome_No:22								
2	!Region_Limit:500,0Kb								
3	!Sample_Type:0								
4	No	Marker1	Marker2	SNP1 ID	SNP2 ID	Distance	|D'|	r2	LOD
5	1	0	1	SNP_A-21,...	SNP_A-20,...	89823	0,09500	0,00764	0,8671
6	2	0	2	SNP_A-21,...	SNP_A-20,...	228013	0,09307	0,00163	0,1585
7	3	0	3	SNP_A-21,...	SNP_A-20,...	1631267	0,69114	0,00829	1,0149
8	4	0	4	SNP_A-21,...	SNP_A-20,...	1631467	0,77396	0,01150	1,5178
9	5	0	5	SNP_A-21,...	SNP_A-21,...	1631723	0,24361	0,00438	0,5080
10	6	0	6	SNP_A-21,...	SNP_A-22,...	2201953	0,00891	0,00005	0,0041
11	7	0	7	SNP_A-21,...	SNP_A-19,...	2203932	0,00891	0,00005	0,0041
12	8	0	8	SNP_A-21,...	SNP_A-21,...	2210468	0,07796	0,00229	0,1888
13	9	0	9	SNP_A-21,...	SNP_A-21,...	2223836	0,02796	0,00046	0,0418
14	10	0	10	SNP_A-21,...	SNP_A-22,...	2482199	0,04832	0,00090	0,0793
15	11	0	11	SNP_A-21,...	SNP_A-21,...	2500275	0,06101	0,00140	0,1285
16	12	0	12	SNP_A-21,...	SNP_A-18,...	3386812	0,07345	0,00119	0,1528
17	13	0	13	SNP_A-21,...	SNP_A-19,...	4760413	0,03308	0,00002	0,0020
18	14	0	14	SNP_A-21,...	SNP_A-22,...	4796341	0,36314	0,01251	1,3947
19	15	0	15	SNP_A-21,...	SNP_A-19,...	4874126	0,32164	0,01264	1,2617
20	16	0	16	SNP_A-21,...	SNP_A-22,...	4874259	0,04558	0,00082	0,0742

<Figure 4-38> Extract pairwise LD calculation result



*	1	2	3	4	5	6
1	!R2_Threshold:0.800					
2	!Method:Simple					
3	!Distance_Limit:500.0,...					
4	!Sample_Type:0					
5	No	Chromoso...	Tagger	Feature ID	Captured	R_square
6	1	22	6	SNP_A-22...	7,9	1,000,0,976
7	2	22	28	SNP_A-19...	29	1,000
8	3	22	37	SNP_A-42...	38	1,000
9	4	22	10	SNP_A-22...	11	0,938
10	5	22	3	SNP_A-20...	4	0,904
11	6	22	30	SNP_A-20...	33	0,904
12	7	22	49	SNP_A-22...	50	0,821
13	8	22	0	SNP_A-21...	singleton	
14	9	22	1	SNP_A-20...	singleton	
15	10	22	2	SNP_A-20...	singleton	
16	11	22	5	SNP_A-21...	singleton	
17	12	22	8	SNP_A-21...	singleton	
18	13	22	12	SNP_A-18...	singleton	
19	14	22	13	SNP_A-19...	singleton	

<Figure 4-39> Extract tagging SNP calculation result

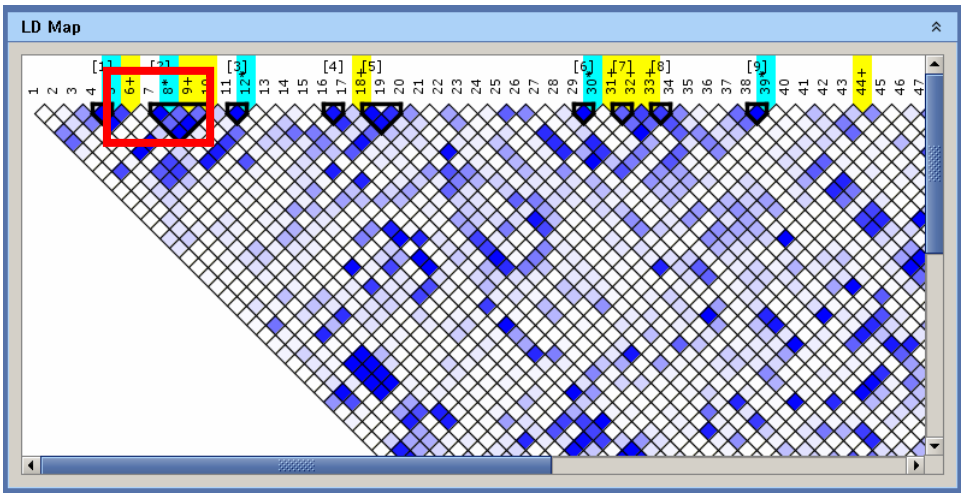


*	1	2	3	4	5	6
1	!Chromos...					
2	!Sample_...					
3	!Block_No...					
4	!Haplotype...					
5	!Multi_Dpri...					
6	No	Haplotype_1	Haplotype_2	Haplotype...	Haplotype...	Crossover_Rate
7	1	22_0_1_1	22_0_2_1	CT	CCCC	0,52426
8	2	22_0_1_1	22_0_2_2	CT	TTCG	0,31098
9	3	22_0_1_1	22_0_2_3	CT	CCTC	0,09883
10	4	22_0_1_1	22_0_2_4	CT	TTCC	0,00549
11	5	22_0_1_2	22_0_2_1	AG	CCCC	0,03631
12	6	22_0_1_2	22_0_2_2	AG	TTCG	0,01864
13	7	22_0_1_2	22_0_2_3	AG	CCTC	0,00000
14	8	22_0_1_2	22_0_2_4	AG	TTCC	0,00000
15	9	22_0_1_3	22_0_2_1	CG	CCCC	0,00000
16	10	22_0_1_3	22_0_2_2	CG	TTCG	0,00549
17	11	22_0_1_3	22_0_2_3	CG	CCTC	0,00000
18	12	22_0_1_3	22_0_2_4	CG	TTCC	0,00000
19	!Chromos...					

<Figure 4-40> Extract haplotype relationships in each LD block

4.6.2. LD Blocking Control and LD Map Visualization Panel

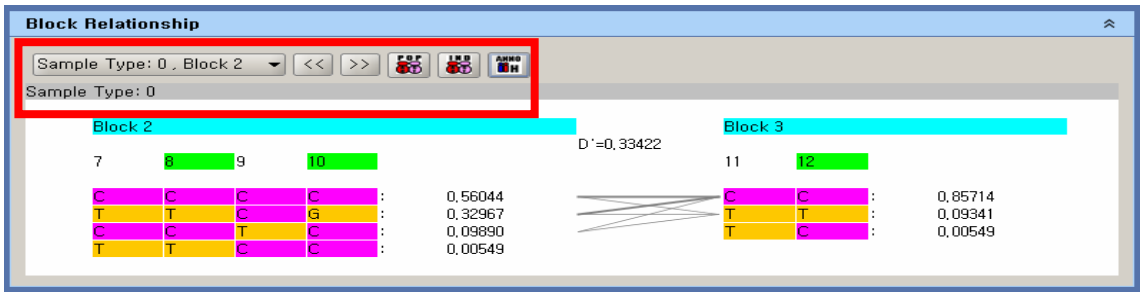
<Figure 4-41> shows degree of linkage disequilibrium of adjacent SNPs and LD blocks. As the values of |D'| or R^2 are closer to 1, the color becomes darker (Red, Blue, and Green). And as they are closer to 0, the color becomes closer to white. Tagging SNP is shown as a light blue square with "*" in the top part of LD map. If a SNP is estimated to be statistically significant in the cross tabulation analysis, it is shown as a yellow square with "+". In order to view haplotype information, haplotype tagging SNP information and haplotype relationship between LD blocks, click the specified block and a figure as in <Figure 4-42> appears in the bottom of LD Map.



<Figure 4-41> LD Map figure

In order to view haplotype information, haplotype tagging SNP information, and relationship between adjacent LD blocks, just click the specified LD block, then <Figure 4-42> shows. Use the button in the upper left of the screen, or [<<] or [>>] to browse other block information.

- Click the IND button to show the estimated haplotype of each individual in a specified block in table format.
- Click the POP button to show the haplotypes and haplotype frequencies in a specified block
- Click the ANNO button to show the annotation information for the SNPs that form haplotype in a specified block in table format.



<Figure 4-42> Relationship between haplotypes in adjacent LD blocks

Chapter 5

Export

5. Export Analysis Result & Biological Annotation

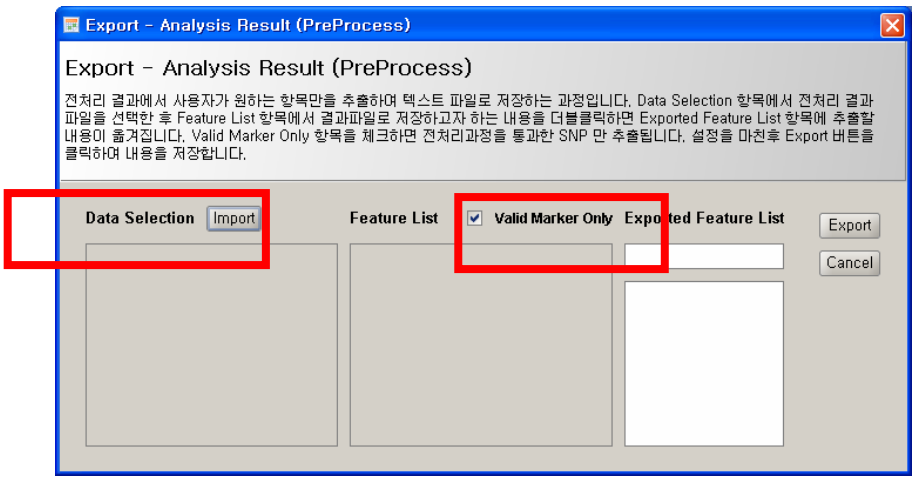
Users can extract and save diverse analysis results and biological annotation information about SNPs.

5.1. Export Analysis Result

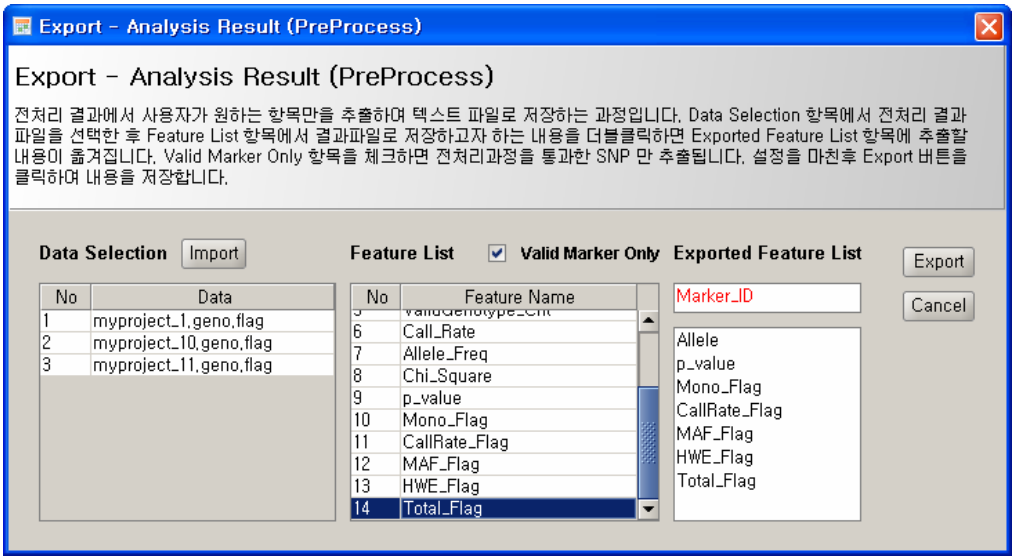
5.1.1. Export PreProcess

Click [Export] > [Analysis Result] > [PreProcess] to show a window where you can extract the preprocessing result as text file as in <Figure 5-1>.

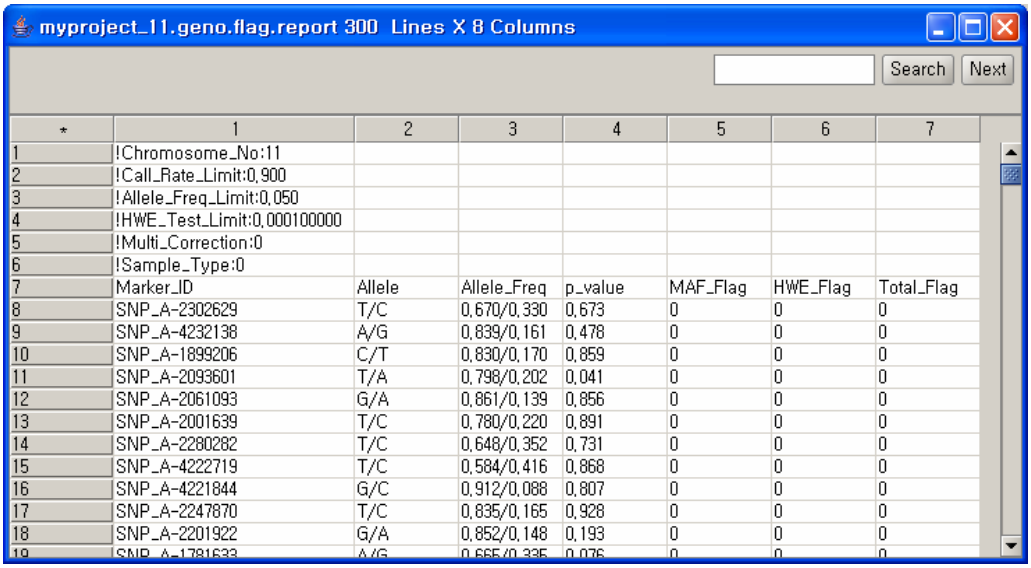
- Click [Import] to show a window where you can select a preprocessing result file (*.flag file) and you can select one or more files using the <CTRL> key.
- If you want to save only the results of the SNPs passing the preprocessing threshold, just check the “Valid Marker Only”.
- The contents of the anaysis results are displayed in “Feature List” as in <Figure 5-2>. Double-click a feature to extract and it will move to the “Exported Feature List”.
- Click [Export] after finishing setting process. The extracted contents are displayed as in <Figure 5-3> and the extracted contents are added in the “Report” tab of project tree.



<Figure 5-1> Extract preprocessing results



<Figure 5-2> Designation of contents to be extracted

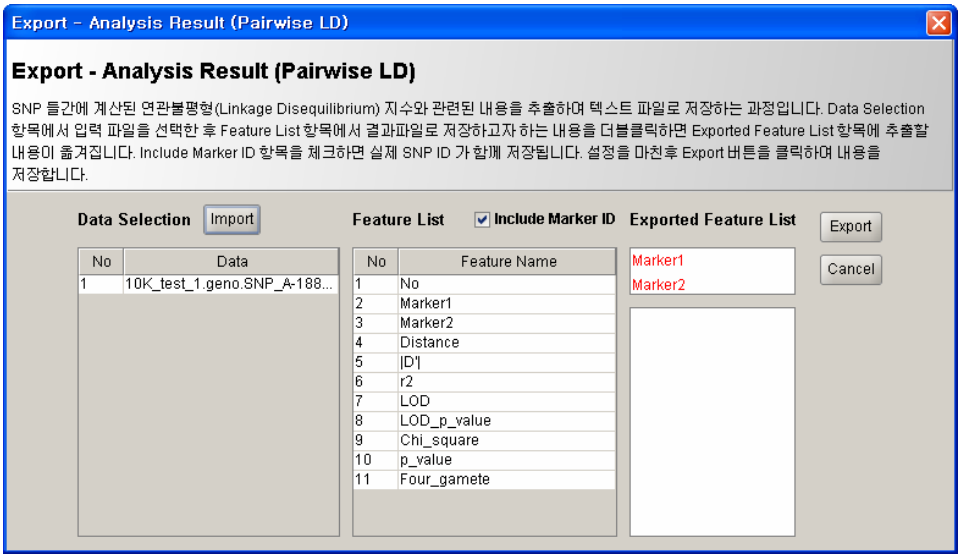


*	1	2	3	4	5	6	7
1	!Chromosome_No:11						
2	!Call_Rate_Limit:0,900						
3	!Allele_Freq_Limit:0,050						
4	!HWE_Test_Limit:0,000100000						
5	!Multi_Correction:0						
6	!Sample_Type:0						
7	Marker_ID	Allele	Allele_Freq	p_value	MAF_Flag	HWE_Flag	Total_Flag
8	SNP_A-2302629	T/C	0,670/0,330	0,673	0	0	0
9	SNP_A-4232138	A/G	0,839/0,161	0,478	0	0	0
10	SNP_A-1899206	C/T	0,830/0,170	0,859	0	0	0
11	SNP_A-2093601	T/A	0,798/0,202	0,041	0	0	0
12	SNP_A-2061093	G/A	0,861/0,139	0,856	0	0	0
13	SNP_A-2001639	T/C	0,780/0,220	0,891	0	0	0
14	SNP_A-2280282	T/C	0,648/0,352	0,731	0	0	0
15	SNP_A-4222719	T/C	0,584/0,416	0,868	0	0	0
16	SNP_A-4221844	G/C	0,912/0,088	0,807	0	0	0
17	SNP_A-2247870	T/C	0,835/0,165	0,928	0	0	0
18	SNP_A-2201922	G/A	0,852/0,148	0,193	0	0	0

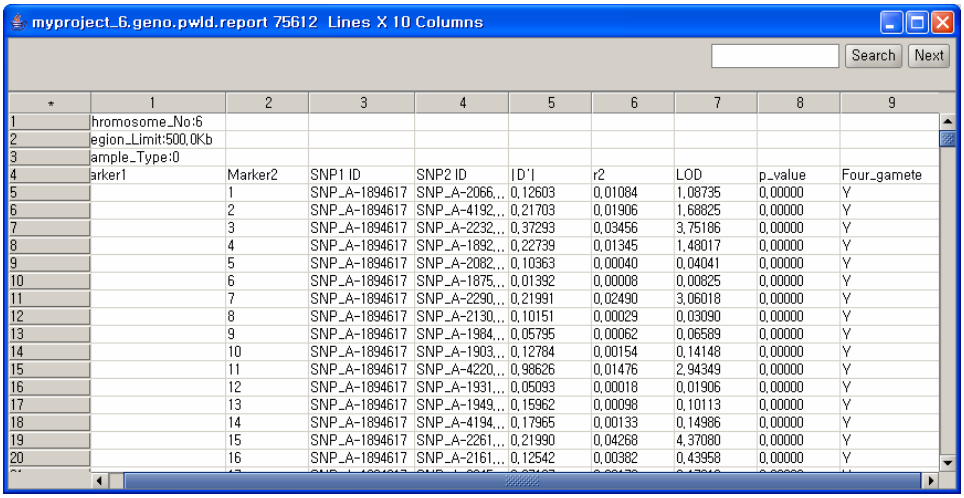
<Figure 5-3> Extracted contents

5.1.2. LD Analysis (Pairwise LD)

Click [Export] > [Analysis Result] > [LD Analysis] > [Pairwise LD] to show a window where you can extract the results of the linkage disequilibrium analysis between SNPs as in <Figure 5-4>. The extraction process is similar to the one described in **5.1.1 Export PreProcess**. If “Include Marker ID” is checked, actual SNP IDs are extracted along with serial numbers. <Figure 5-5> shows the extracted result.



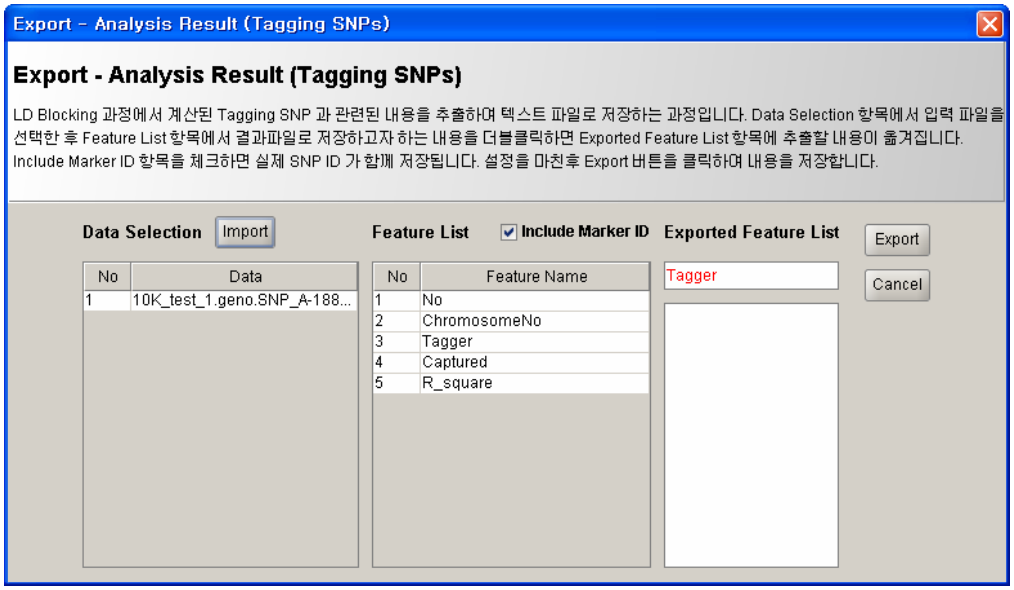
<Figure 5-4> Designation of contents to be extracted



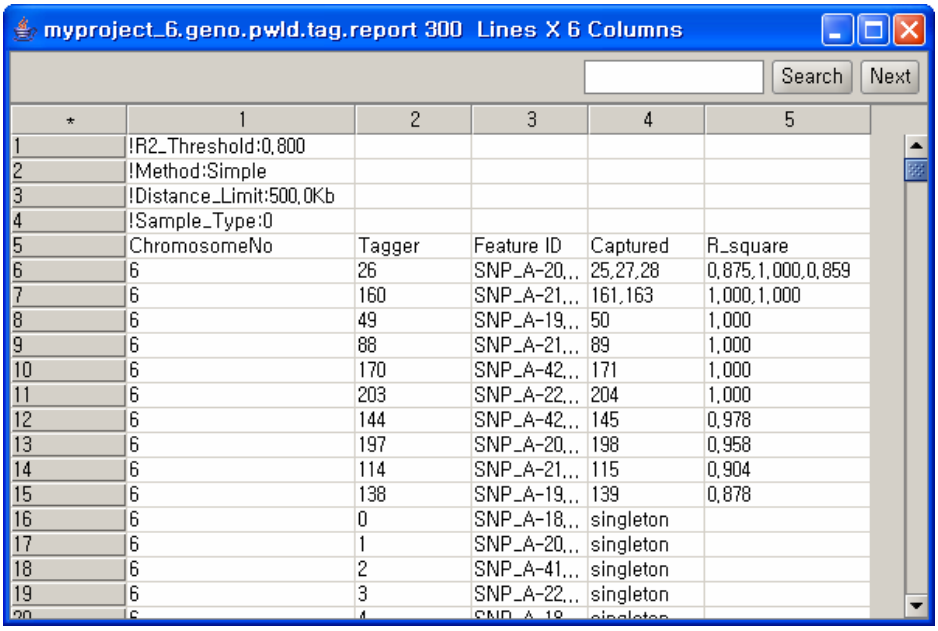
<Figure 5-5> Extracted contents

5.1.3. LD Analysis (Tagging SNPs)

Click [Export] > [Analysis Result] > [LD Analysis] > [Tagging SNPs] to show a window where you can extract the tagging SNPs. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** If “Include Marker ID” is checked, actual SNP IDs are extracted along with serial numbers. <Figure 5-7> shows the extracted result.



<Figure 5-6> Designation of contents to be extracted

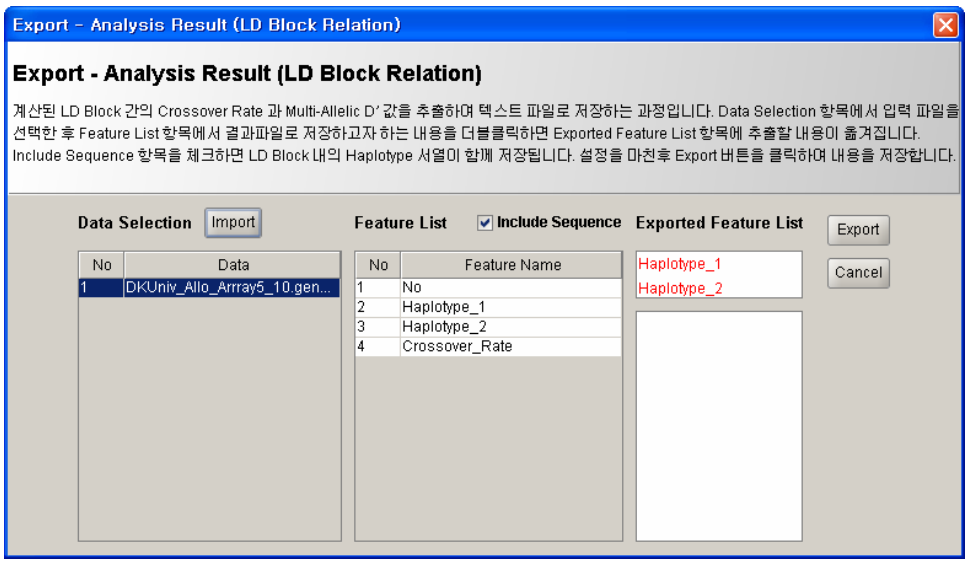


*	1	2	3	4	5
1	!R2_Threshold:0.800				
2	!Method:Simple				
3	!Distance_Limit:500.0Kb				
4	!Sample_Type:0				
5	ChromosomeNo	Tagger	Feature ID	Captured	R_square
6	6	26	SNP_A-20...	25,27,28	0.875,1.000,0.859
7	6	160	SNP_A-21...	161,163	1.000,1.000
8	6	49	SNP_A-19...	50	1.000
9	6	88	SNP_A-21...	89	1.000
10	6	170	SNP_A-42...	171	1.000
11	6	203	SNP_A-22...	204	1.000
12	6	144	SNP_A-42...	145	0.978
13	6	197	SNP_A-20...	198	0.958
14	6	114	SNP_A-21...	115	0.904
15	6	138	SNP_A-19...	139	0.878
16	6	0	SNP_A-18...	singleton	
17	6	1	SNP_A-20...	singleton	
18	6	2	SNP_A-41...	singleton	
19	6	3	SNP_A-22...	singleton	

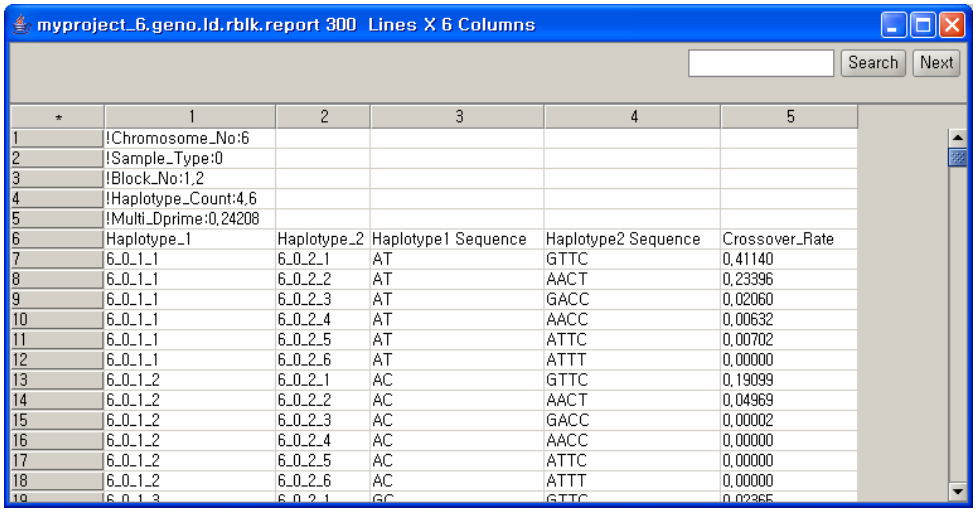
<Figure 5-7> Extracted contents

5.1.4. LD Analysis (LD Block Relationship)

Click [Export] > [Analysis Result] > [LD Analysis] > [LD Block Relationship] to show a window where you can extract the haplotype relationships between adjacent LD blocks as in <Figure 5-8>. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** If “Include Sequence” is checked, the haplotype IDs and haplotype sequences are both extracted. <Figure 5-9> shows the extracted contents.



<Figure 5-8> Designation of contents to be extracted

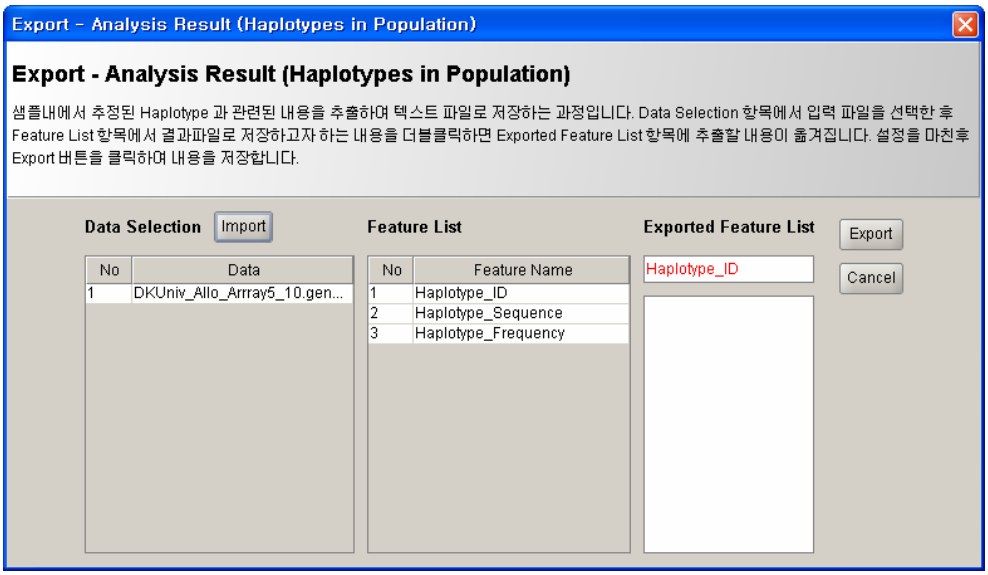


*	1	2	3	4	5
1	!Chromosome_No:6				
2	!Sample_Type:0				
3	!Block_No:1,2				
4	!Haplotype_Count:4,6				
5	!Multi_Dprime:0,24208				
6	Haplotype_1	Haplotype_2	Haplotype1 Sequence	Haplotype2 Sequence	Crossover_Rate
7	6_0_1_1	6_0_2_1	AT	GTTC	0,41140
8	6_0_1_1	6_0_2_2	AT	AACT	0,23396
9	6_0_1_1	6_0_2_3	AT	GACC	0,02060
10	6_0_1_1	6_0_2_4	AT	AACC	0,00632
11	6_0_1_1	6_0_2_5	AT	ATTC	0,00702
12	6_0_1_1	6_0_2_6	AT	ATTT	0,00000
13	6_0_1_2	6_0_2_1	AC	GTTC	0,19099
14	6_0_1_2	6_0_2_2	AC	AACT	0,04969
15	6_0_1_2	6_0_2_3	AC	GACC	0,00002
16	6_0_1_2	6_0_2_4	AC	AACC	0,00000
17	6_0_1_2	6_0_2_5	AC	ATTC	0,00000
18	6_0_1_2	6_0_2_6	AC	ATTT	0,00000
19	6_0_1_3	6_0_2_1	GC	GTTC	0,02365

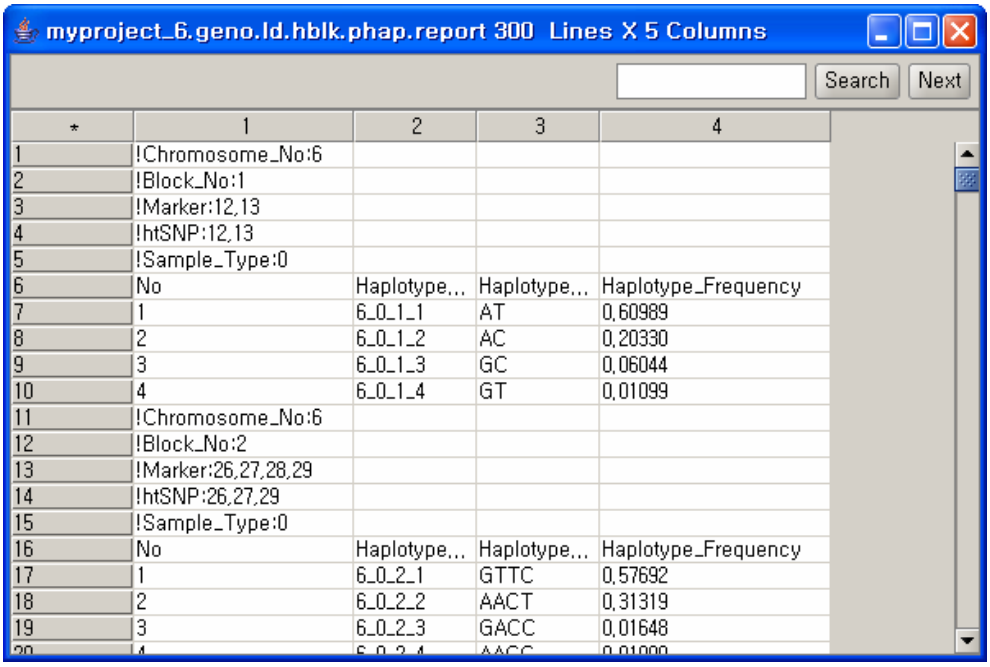
<Figure 5-9> Extracted contents

5.1.5. LD Analysis (Haplotypes in Population)

Click [Export] > [Analysis Result] > [LD Analysis] > [Haplotypes in Population] to show a window where you can extract the haplotypes and theris frequencies in samples as in <Figure 5-10>. The extraction process is similar to the one described in **5.1.1 Export PreProcess**. <Figure 5-11> shows the extracted contents.



<Figure 5-10> Designation of contents to be extracted

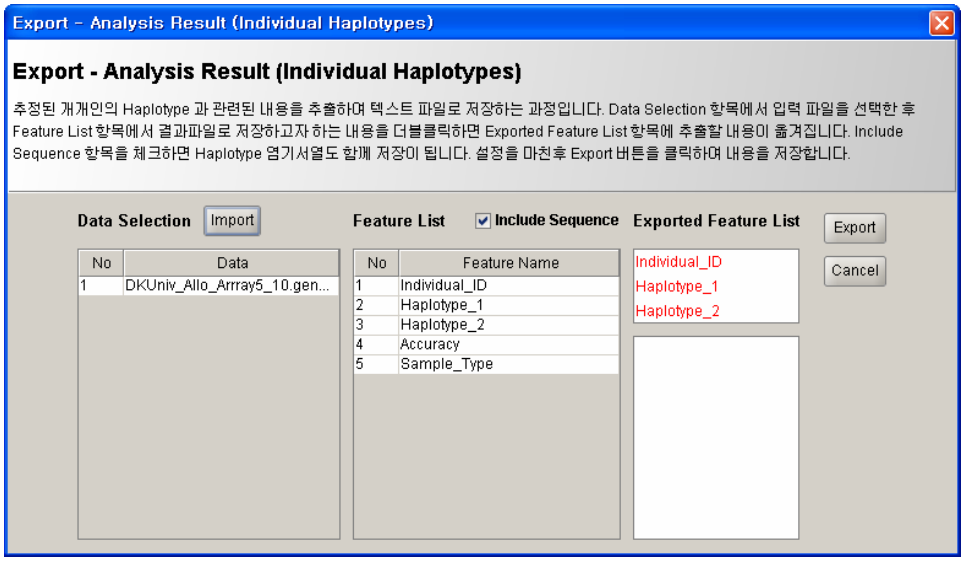


<Figure 5-11> Extracted contents

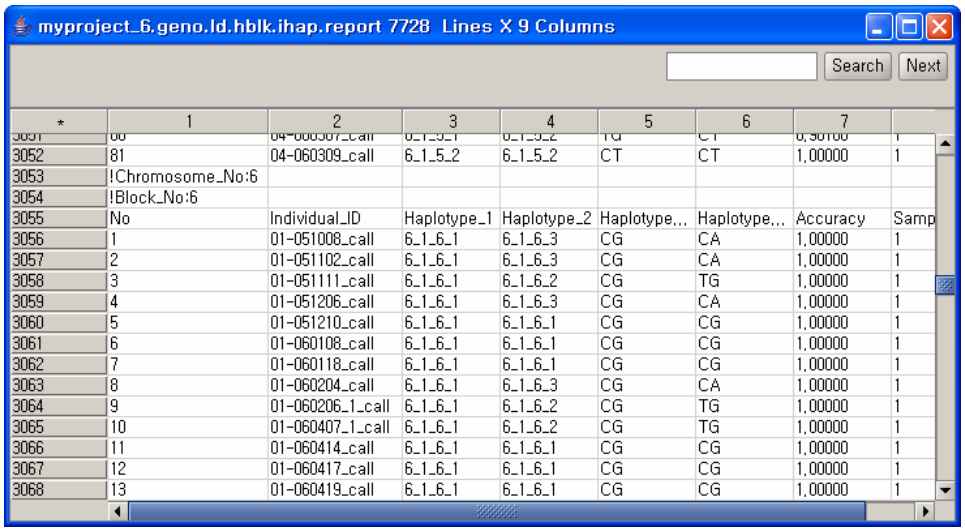
5.1.6. LD Analysis (Individual Haplotype)

Click [Export] > [Analysis Result] > [LD Analysis] > [Individual Haplotype] to show a window where you can extract the individual haplotype set as in <Figure 5-12> The extraction method is similar to the method described in **5.1.1 Export PreProcess.** <Figure 5-13> shows the

extracted contents.



<Figure 5-12> Designation of contents to be extracted

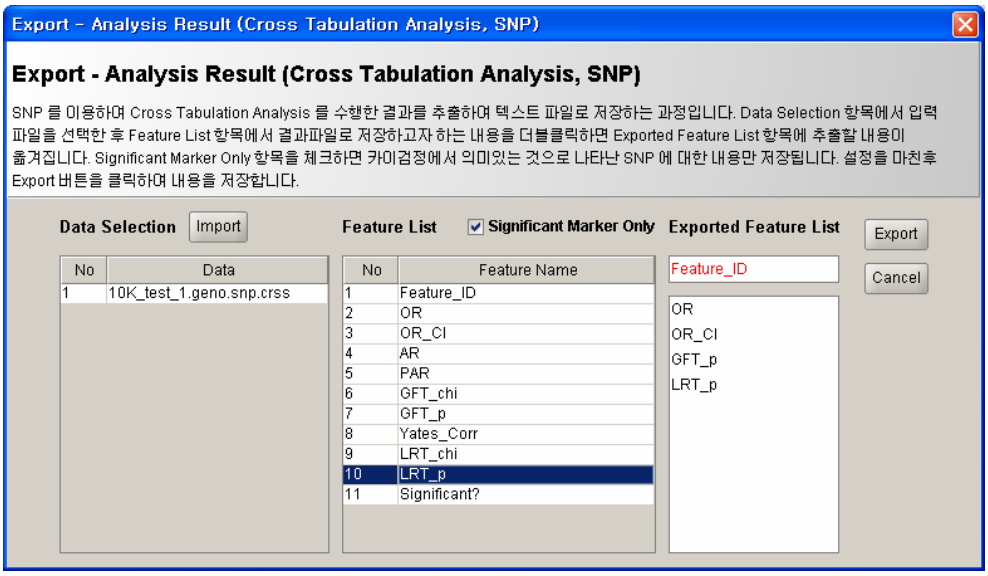


<Figure 5-13> Extracted contents

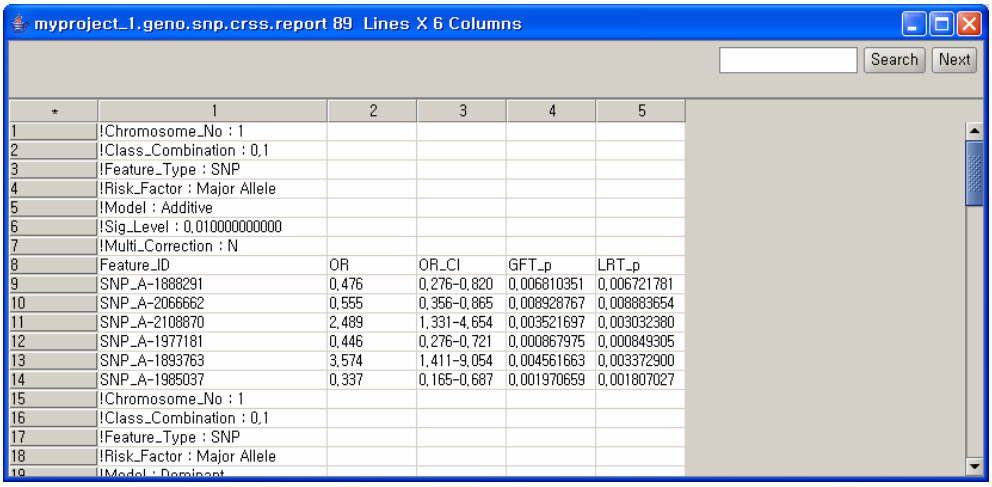
5.1.7. Cross Tabulation Analysis (SNP)

Click [Export] > [Analysis Result] > [Cross Tabulation Analysis(SNP)] to show a window where you can extract the analyzed result of SNPs as in <Figure 5-14>. If “Significant Marker Only” is checked, only the SNPs that pass the significance level. The extraction process is similar to the one described in **5.1.1 Export PreProcess**. <Figure 5-15> shows the extracted

contents.



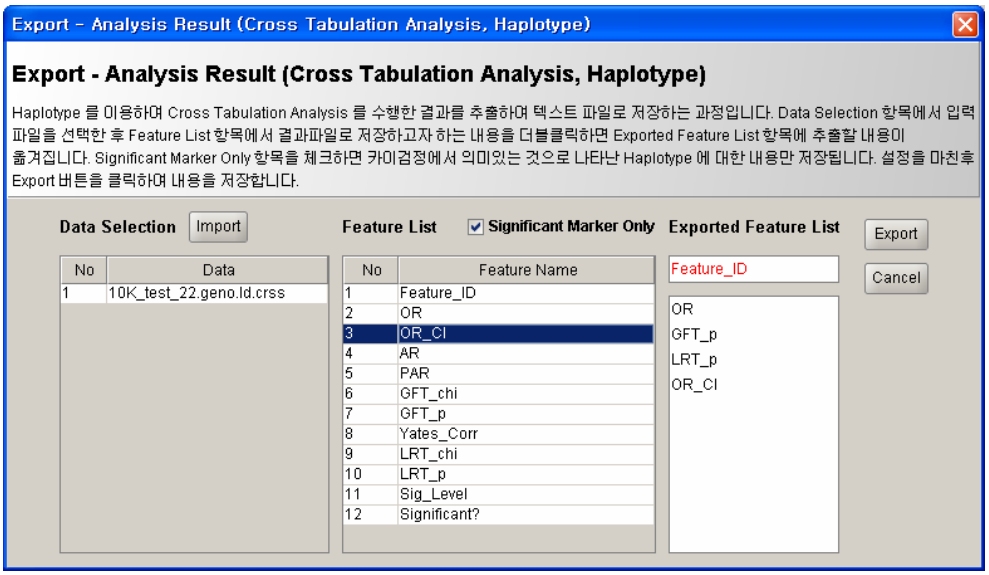
<Figure 5-14> Designation of contents to be extracted



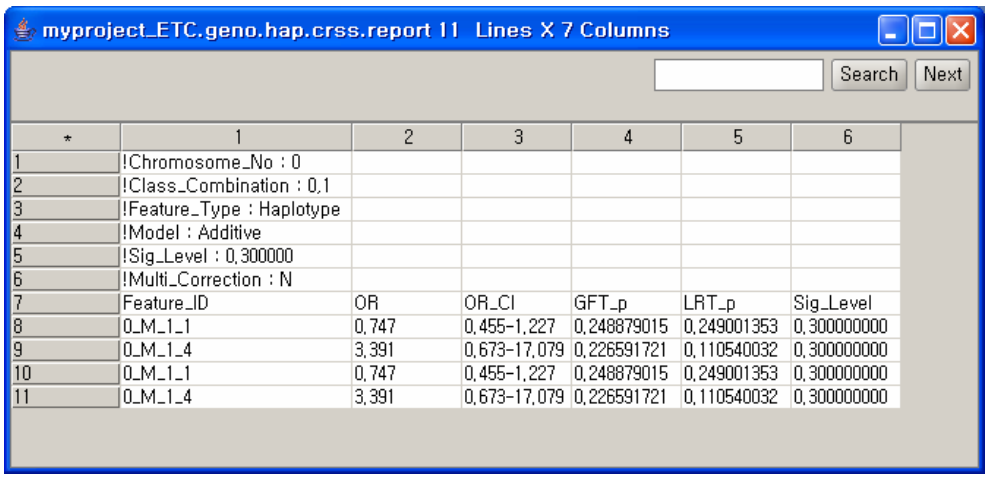
<Figure 5-15> Extracted contents

5.1.8. Cross Tabulation Analysis (Haplotype)

Click [Export]→[Analysis Result]→[Cross Tabulation Analysis(Haplotype)] to show a window where you can extract the analysis result as in <Figure 5-16>. If “Significant Marker Only” is checked, it extracts only the haplotypes that pass the significance level set when performing analysis. The extraction method is similar to the method described in **5.1.1 Export PreProcess.** <Figure 5-17> shows the extracted contents.



<Figure 5-16> Designation of contents to be extracted

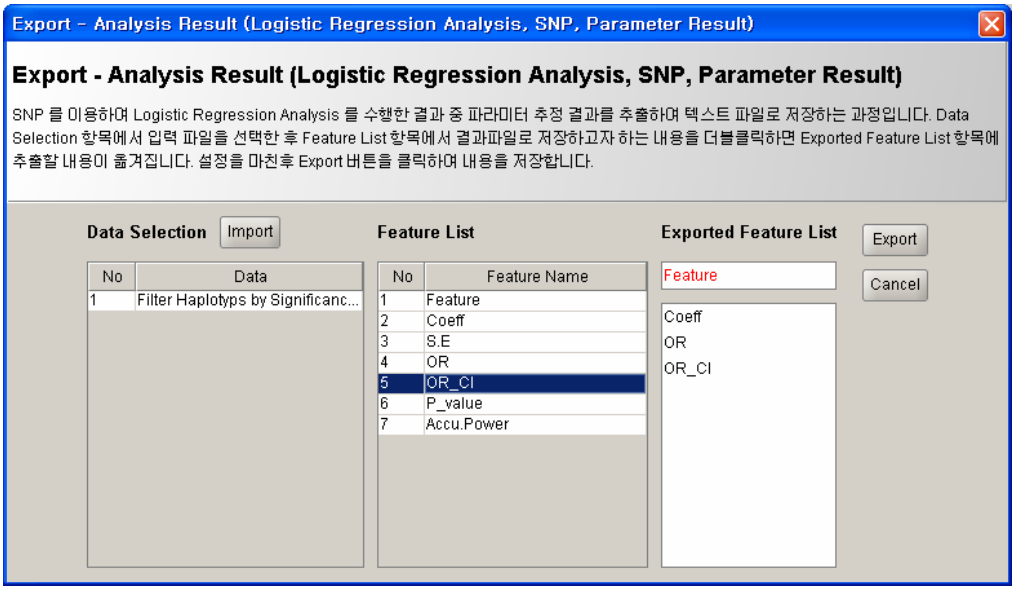


*	1	2	3	4	5	6
1	!Chromosome_No : 0					
2	!Class_Combination : 0,1					
3	!Feature_Type : Haplotype					
4	!Model : Additive					
5	!Sig_Level : 0,300000					
6	!Multi_Correction : N					
7	Feature_ID	OR	OR_CI	GFT_p	LRT_p	Sig_Level
8	0_M_1_1	0,747	0,455-1,227	0,248879015	0,249001353	0,300000000
9	0_M_1_4	3,391	0,673-17,079	0,226591721	0,110540032	0,300000000
10	0_M_1_1	0,747	0,455-1,227	0,248879015	0,249001353	0,300000000
11	0_M_1_4	3,391	0,673-17,079	0,226591721	0,110540032	0,300000000

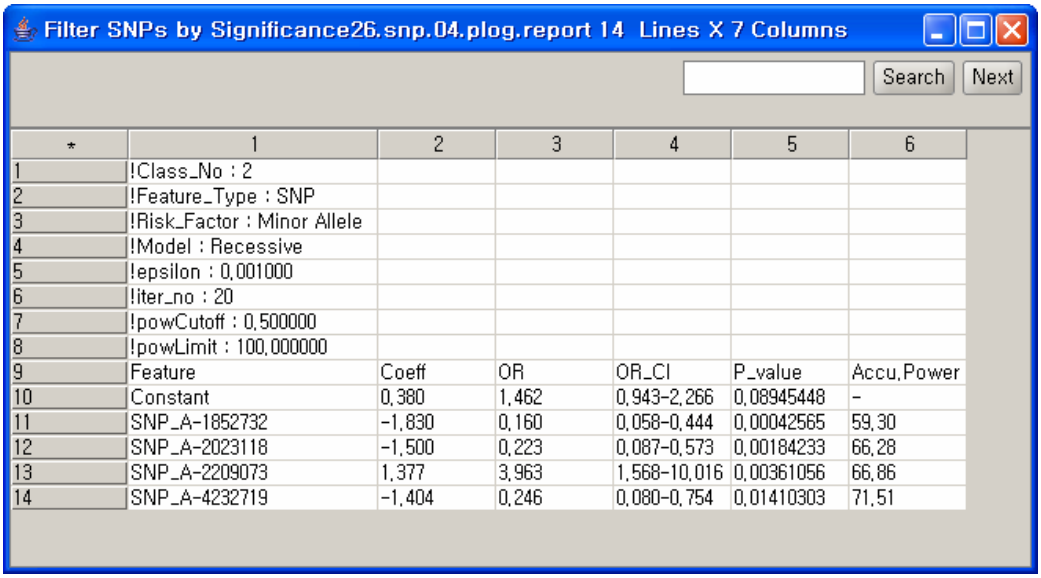
<Figure 5-17> Extracted contents

5.1.9. Logistic Regression Analysis (SNP, Parameter Estimation)

Click [Export] > [Analysis Result] > [Logistic Regression Analysis(SNP)] > [Parameter Estimation] to show a window where you can extract the estimatied coefficient of each SNP in the logistic regression model as in <Figure 5-18>. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** <Figure 5-19> shows the extracted contents.



<Figure 5-18> Designation of contents to be extracted

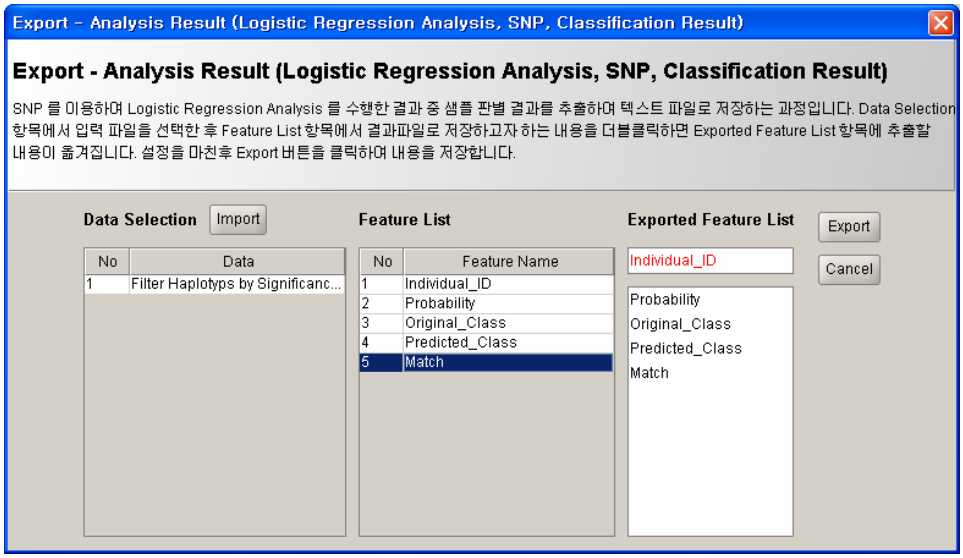


*	1	2	3	4	5	6
1	!Class_No : 2					
2	!Feature_Type : SNP					
3	!Risk_Factor : Minor Allele					
4	!Model : Recessive					
5	!epsilon : 0,001000					
6	!iter_no : 20					
7	!powCutoff : 0,500000					
8	!powLimit : 100,000000					
9	Feature	Coeff	OR	OR_CI	P_value	Accu.Power
10	Constant	0,380	1,462	0,943-2,266	0,08945448	-
11	SNP_A-1852732	-1,830	0,160	0,058-0,444	0,00042565	59,30
12	SNP_A-2023118	-1,500	0,223	0,087-0,573	0,00184233	66,28
13	SNP_A-2209073	1,377	3,963	1,568-10,016	0,00361056	66,86
14	SNP_A-4232719	-1,404	0,246	0,080-0,754	0,01410303	71,51

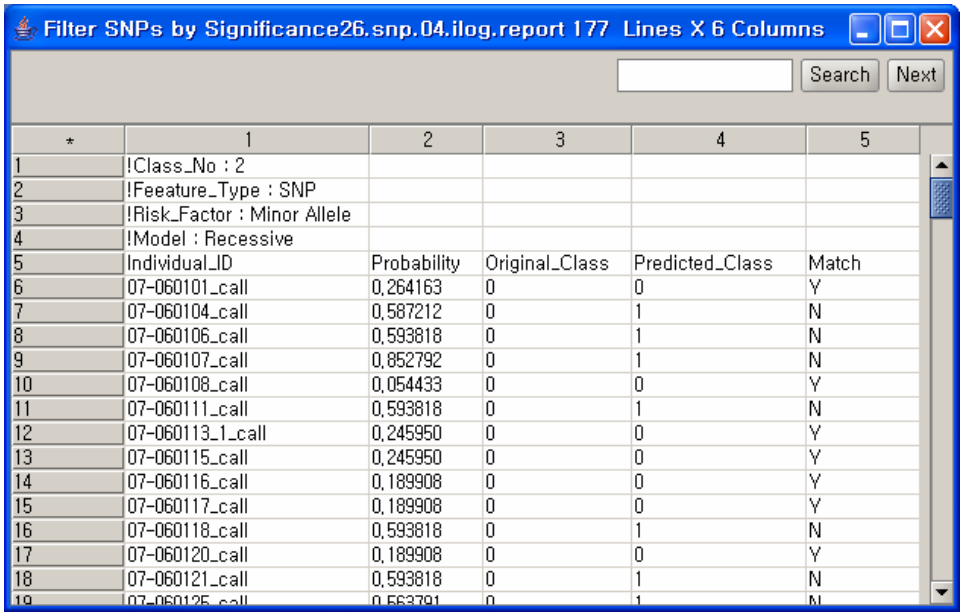
<Figure 5-19> Extracted contents

5.1.10. Logistic Regression Analysis (SNP, Classification Result)

Click [Export] > [Analysis Result] > [Logistic Regression Analysis(SNP)] > [Classification Result] to show a window where you can extract the classification result using SNPs as in <Figure 5-20>. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** <Figure 5-21> show the extracted contents.



<Figure 5-20> Designation of contents to be extracted

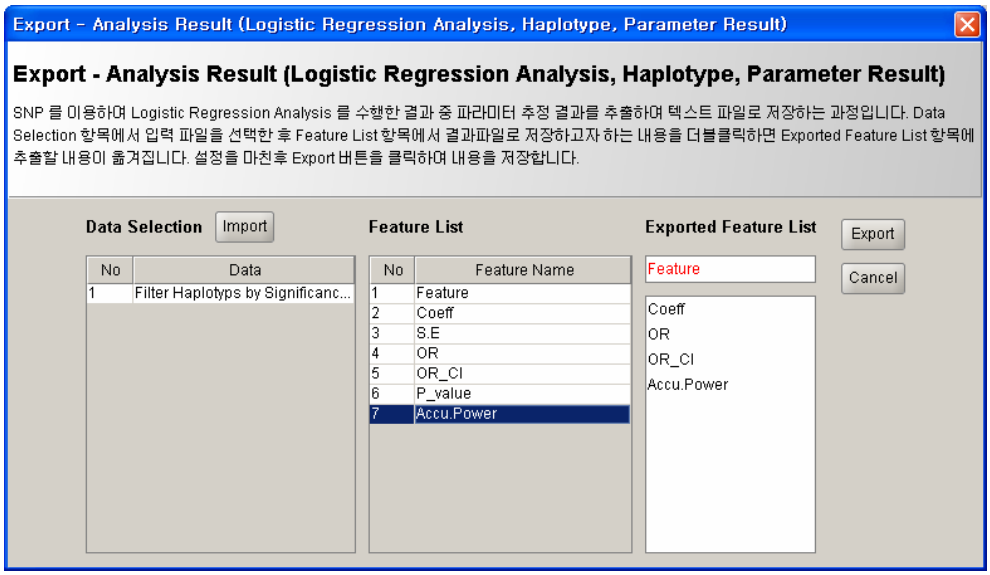


*	1	2	3	4	5
1	!Class_No : 2				
2	!Feeature_Type : SNP				
3	!Risk_Factor : Minor Allele				
4	!Model : Recessive				
5	Individual_ID	Probability	Original_Class	Predicted_Class	Match
6	07-060101_call	0,264163	0	0	Y
7	07-060104_call	0,587212	0	1	N
8	07-060106_call	0,593818	0	1	N
9	07-060107_call	0,852792	0	1	N
10	07-060108_call	0,054433	0	0	Y
11	07-060111_call	0,593818	0	1	N
12	07-060113_1_call	0,245950	0	0	Y
13	07-060115_call	0,245950	0	0	Y
14	07-060116_call	0,189908	0	0	Y
15	07-060117_call	0,189908	0	0	Y
16	07-060118_call	0,593818	0	1	N
17	07-060120_call	0,189908	0	0	Y
18	07-060121_call	0,593818	0	1	N
19	07-060125_call	0,563791	0	1	N

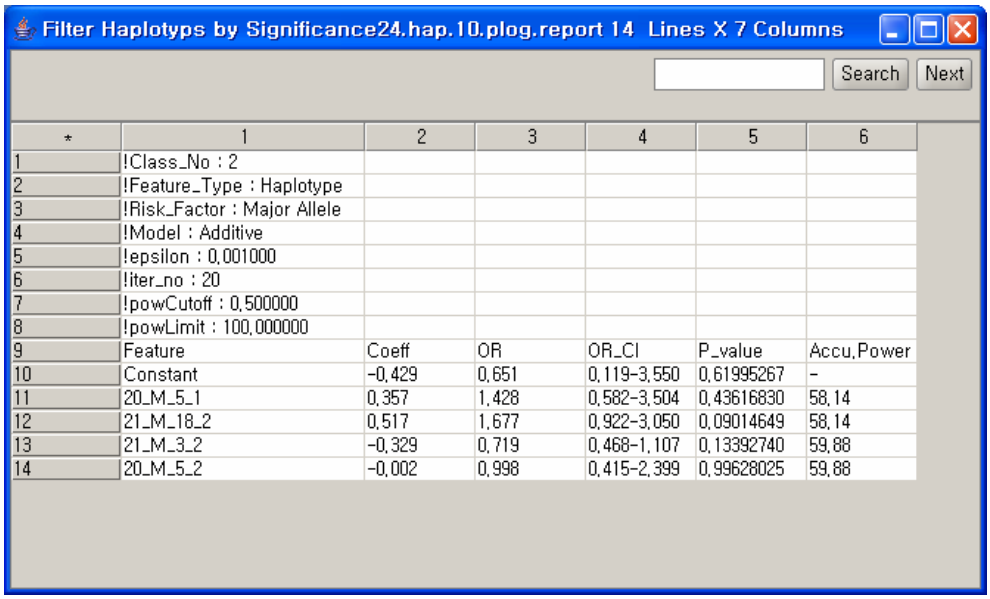
<Figure 5-21> Extracted contents

5.1.11. Logistic Regression Analysis (Haplotype, Parameter Estimation)

Click [Export] > [Analysis Result] > [Logistic Regression Analysis(Haplotype)] > [Parameter Estimation] to show a window where you can extract the estimated coefficient of each haplotype in the logistic model as in <Figure 5-22>. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** <Figure 5-23> shows the extracted contents.



<Figure 5-22> Designation of contents to be extracted

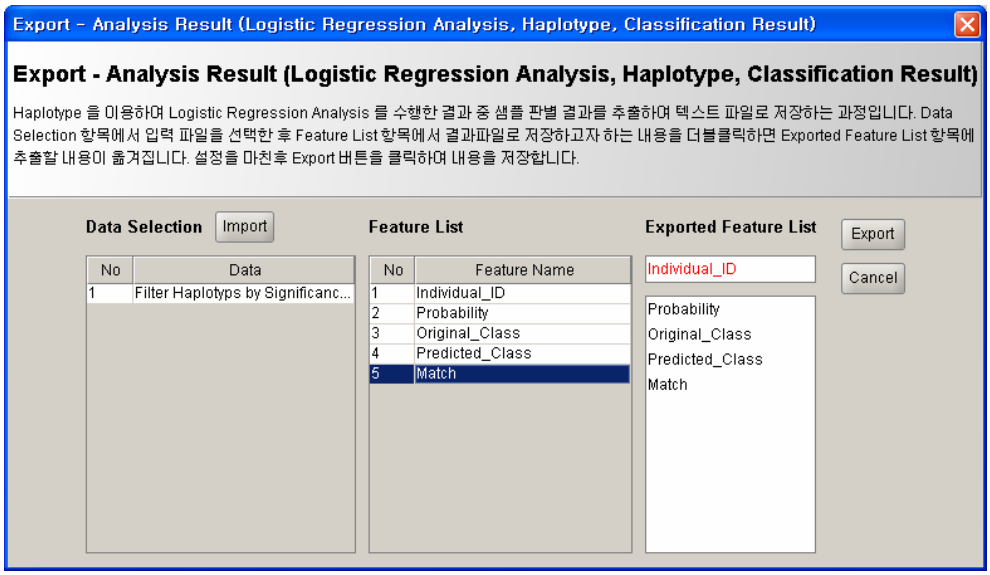


*	1	2	3	4	5	6
1	!Class_No : 2					
2	!Feature_Type : Haplotype					
3	!Risk_Factor : Major Allele					
4	!Model : Additive					
5	!epsilon : 0,001000					
6	!iter_no : 20					
7	!powCutoff : 0,500000					
8	!powLimit : 100,000000					
9	Feature	Coeff	OR	OR_CI	P_value	Accu,Power
10	Constant	-0,429	0,651	0,119-3,550	0,61995267	-
11	20_M_5_1	0,357	1,428	0,582-3,504	0,43616830	58,14
12	21_M_18_2	0,517	1,677	0,922-3,050	0,09014649	58,14
13	21_M_3_2	-0,329	0,719	0,468-1,107	0,13392740	59,88
14	20_M_5_2	-0,002	0,998	0,415-2,399	0,99628025	59,88

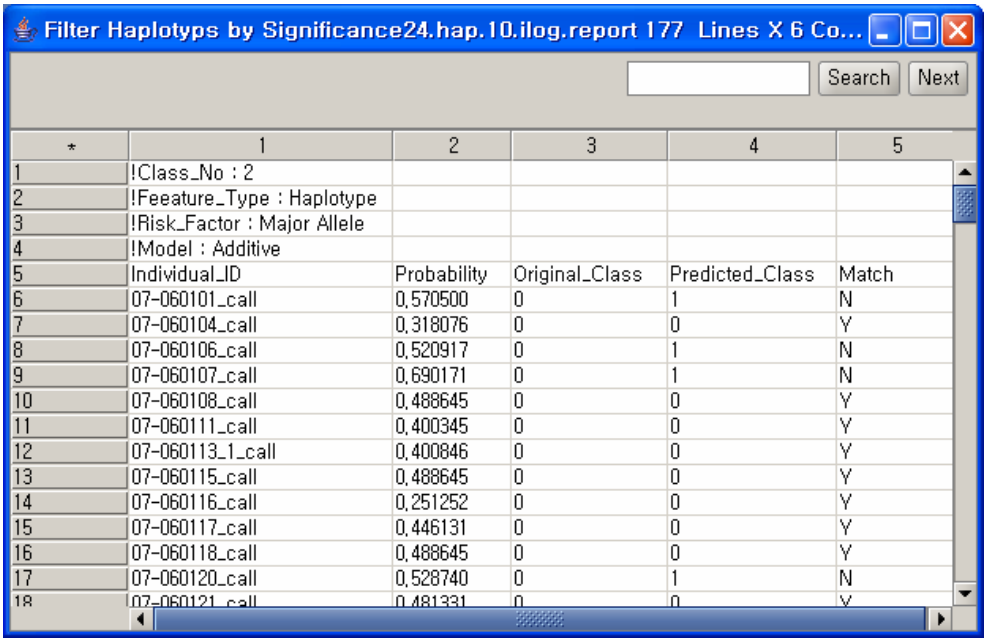
<Figure 5-23> extracted contents

5.1.12. Logistic Regression Analysis (Haplotype, Classification Result)

Click [Export] > [Analysis Result] > [Logistic Regression Analysis (Haplotype)] > [Classification Result] to show a window where you can extract the classification result using haplotypes as in <Figure 5-24>. The extraction process is similar to the one described in **5.1.1 Export PreProcess.** <Figure 5-25> shows the extracted contents.



<Figure 5-24> Designation of contents to be extracted



<Figure 5-25> Extracted contents

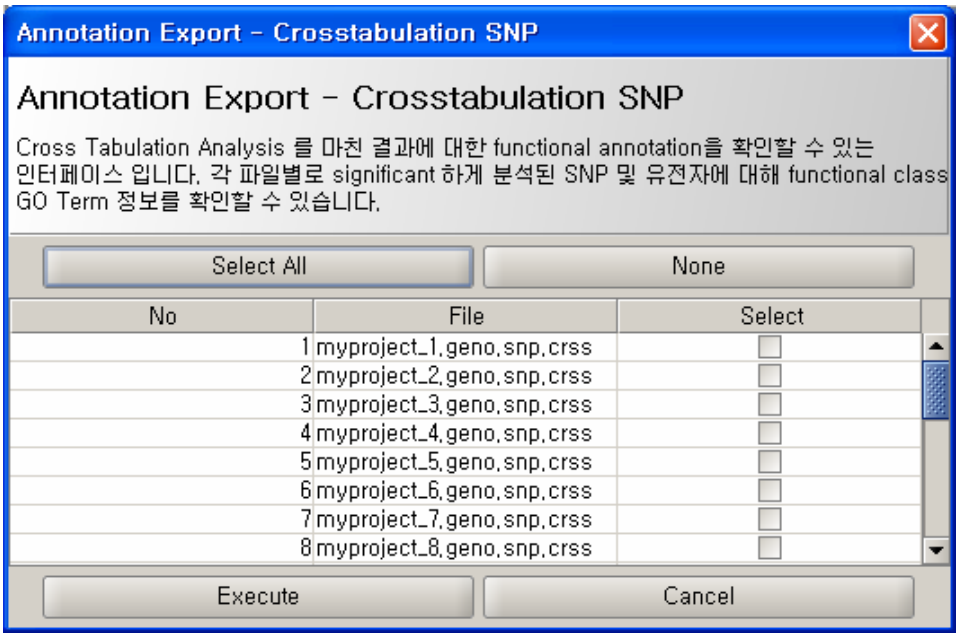
5.2. Export Annotation

5.2.1. Export Annotation of Cross Tabulation Analysis (SNP)

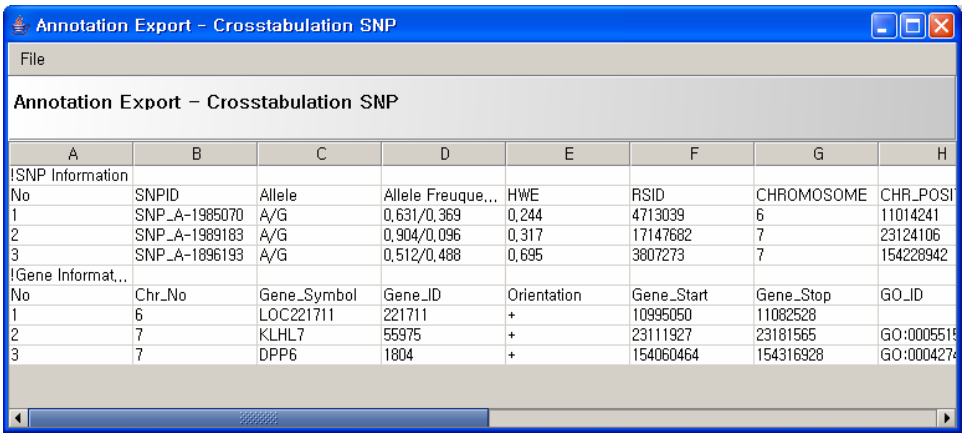
Click [Export] > [Annotation (Cross Tabulation Analysis, SNP)] to show a window where you can extract the annotation information of SNPs showing statistically significant difference in 2

classes as in <Figure 5-26>. <Figure 5-27> shows the results fo the extracted contents. Descriptions of extracted annotation information are the following:

- Annotation information related to SNP
  - SNP ID
  - Allele
  - Allele Frequency
  - HWE: p-value of Hardy-Weinberg Equilibrium Test
  - Chromosome: chromosome number in which SNP is located
  - Chr_Position: physical position of SNP in chromosome
  - RS_ID: dbSNP #rs of SNP
  - Contig: contig number in which SNP is located
  - Contig_Position: physical position of SNP in contig
  - Gene_ID: NCBI gene ID of the gene in which SNP is located
  - Gene_Symbol: gene symbol of the gene in which SNP is located
  - mRNA: transcript ID
  - Product: protein ID
  - Func_Position: functional class of SNP
    - Type: Non-Synonymous, Synonymous, Intron, Locus Region, Up/Down Stream
- Annotation Information related to gene
  - Chr_No: chromosome number in which gene is located
  - Gene_Symbol
  - Gene_ID: NCBI Gene ID
  - Orientation: orientation of the gene
  - Gene_Start: start position of the gene in the specified chromosome
  - Gene_Stop: stop position of the gene in the specified chromosome
  - GO_ID: gene ontology ID
  - GO_Term: GO term
  - Category: GO category



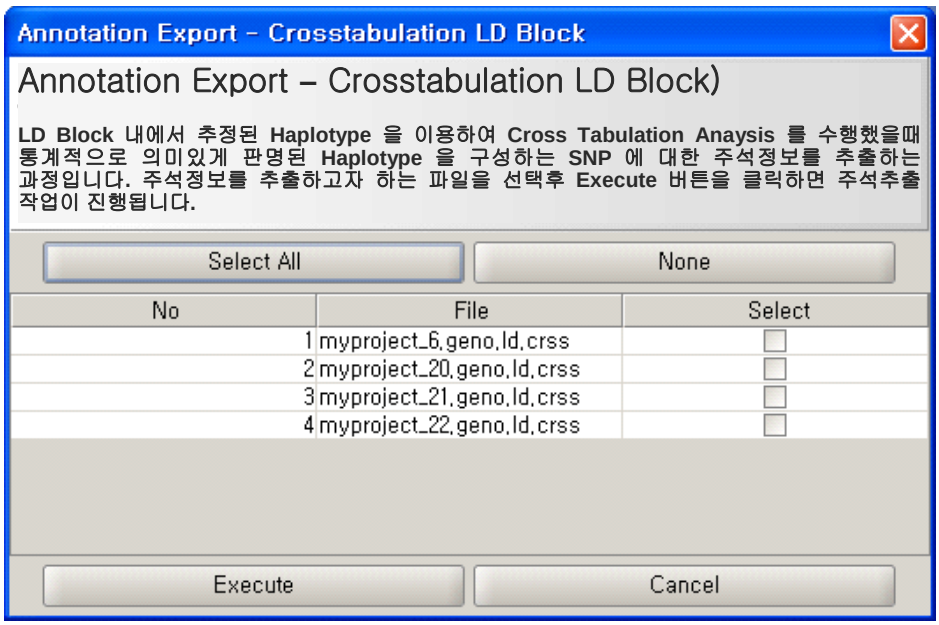
<Figure 5-26> Designation of contents to be extracted



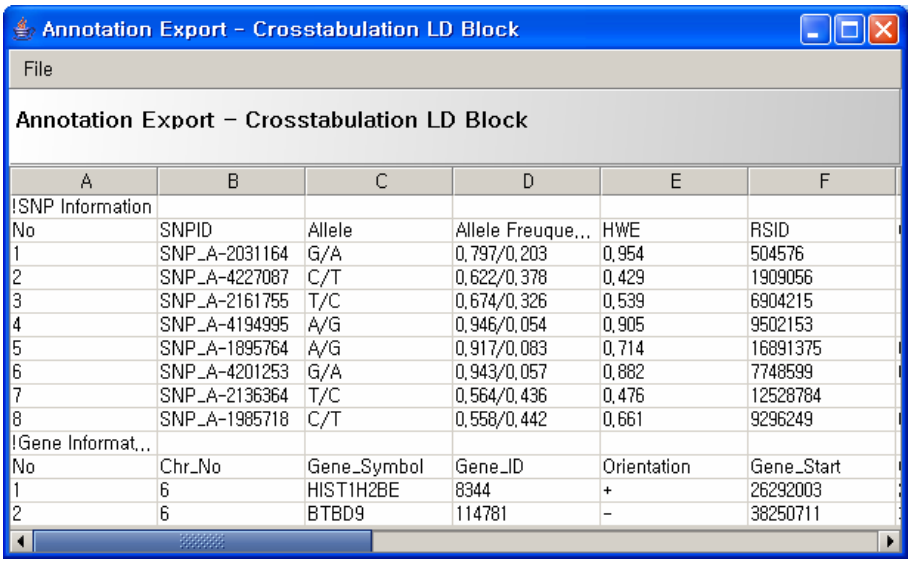
<Figure 5-27> Extracted Biological Annotation Information

5.2.2. Export Annotation of Cross Tabulation Analysis (LD Block)

Click [Export] > [Annotation (Cross Tabulation Analysis, LD Block)] to show a window where you can extract the biological annotation information of SNPs contained in the haplotype that is estimated in the LD block. Click [Execute] after selecting files from which you want to extract annotation information as in <Figure 5-29>. The details of the extracted annotation information are the same as in 5.2.1 Export Annotation of Cross Tabulation Analysis (SNP).



<Figure 5-28> Designation of contents to be extracted



A	B	C	D	E	F
!SNP Information					
No	SNPID	Allele	Allele Freuque...	HWE	RSID
1	SNP_A-2031164	G/A	0.797/0.203	0.954	504576
2	SNP_A-4227087	C/T	0.622/0.378	0.429	1909056
3	SNP_A-2161755	T/C	0.674/0.326	0.539	6904215
4	SNP_A-4194995	A/G	0.946/0.054	0.905	9502153
5	SNP_A-1895764	A/G	0.917/0.083	0.714	16891375
6	SNP_A-4201253	G/A	0.943/0.057	0.882	7748599
7	SNP_A-2136364	T/C	0.564/0.436	0.476	12528784
8	SNP_A-1985718	C/T	0.558/0.442	0.661	9296249
!Gene Informat...					
No	Chr_No	Gene_Symbol	Gene_ID	Orientation	Gene_Start
1	6	HIST1H2BE	8344	+	26292003
2	6	BTBD9	114781	-	38250711

<Figure 5-29> Extracted bological annotation information

Chapter 6

Filter / Data Transformation / Statistics

6. Filter / Data / Transformation / Statistics

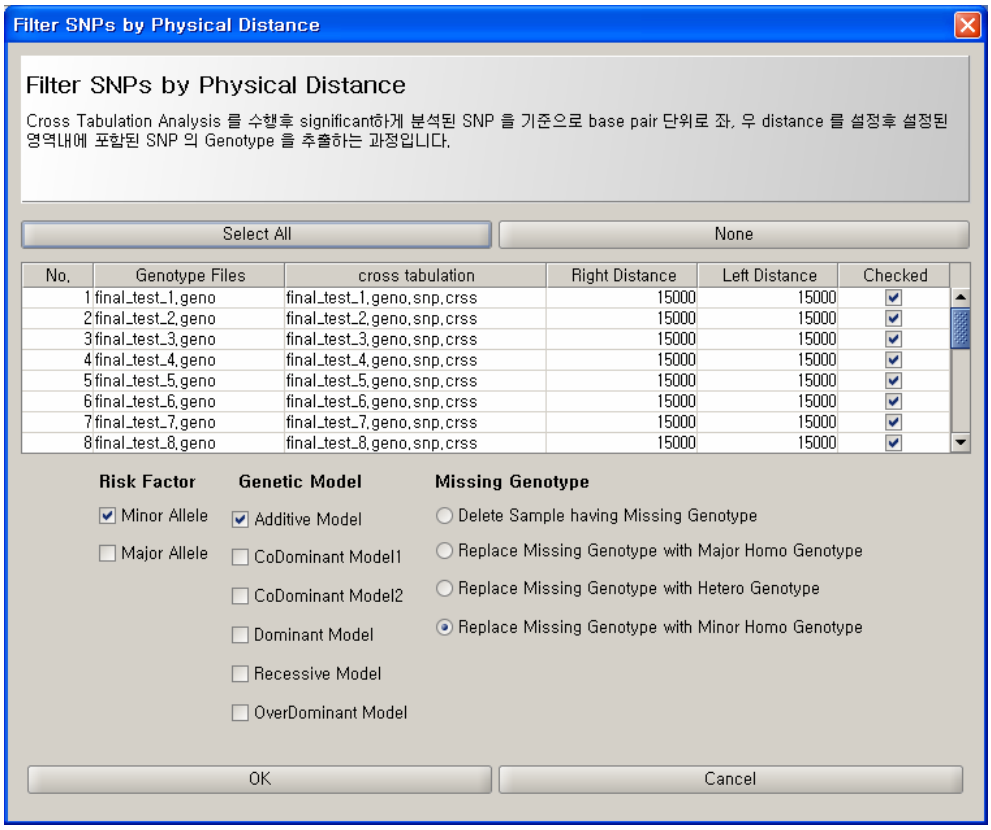
You can filter or transform data required for performing cross tabulation analysis, logistic regression analysis, and LD analysis. The result data from one analysis is needed to be filtered or transformed for the other analysis because all the analyses are computationally connected. For example, the statistically significant SNPs from cross tabulation analysis can be used for logistic regression analysis by transforming genotypes into numerical values. Significant haplotypes from cross tabulation analysis also can be used in logistic regression after data transformation process. If you want to analyze the SNPs contained only in a specific gene, data filintering process can be implemented.

6.1. Filter SNP Data

6.1.1. Filter SNPs by Physical Distance

It is possible to filter SNPs that are adjacent to the statistically significant SNPs by specfing the left and right boundary on the chromosome. Click [Filter] > [Filter SNPs by Physical Distance] to show the setting window as in <Figure 6-1>. The result files from cross tabulation analysis (file extension is *.crss) and genotype files (extension is *.geno) are listed. Enter the proper numbers in the “Right Distance” and “Left Distance”. Click [OK] after setting "Risk Factor, "Genetic Model", and “Missing Genotype" to start the filtering process.

- Risk Factor: Risk Factor set in Cross Tabulation Analysis
  - Minor Allele / Major Allele
- Genetic Model: Analyzed model set in Cross Tabulation Analysis
  - Additive / Codominant1 / Codominant2 / Dominant / Recessive / Overdominant
- Missing Genotype: Missing genotype processing method
  - Reserve Missing Genotype
  - Replace Missing Genotype with Major Homo Genotype
  - Replace Missing Genotype with Hetero Genotype
  - Replace Missing Genotype with Minor Homo Genotype



<Figure 6-1> SNP filtering by specifying distances from left to right

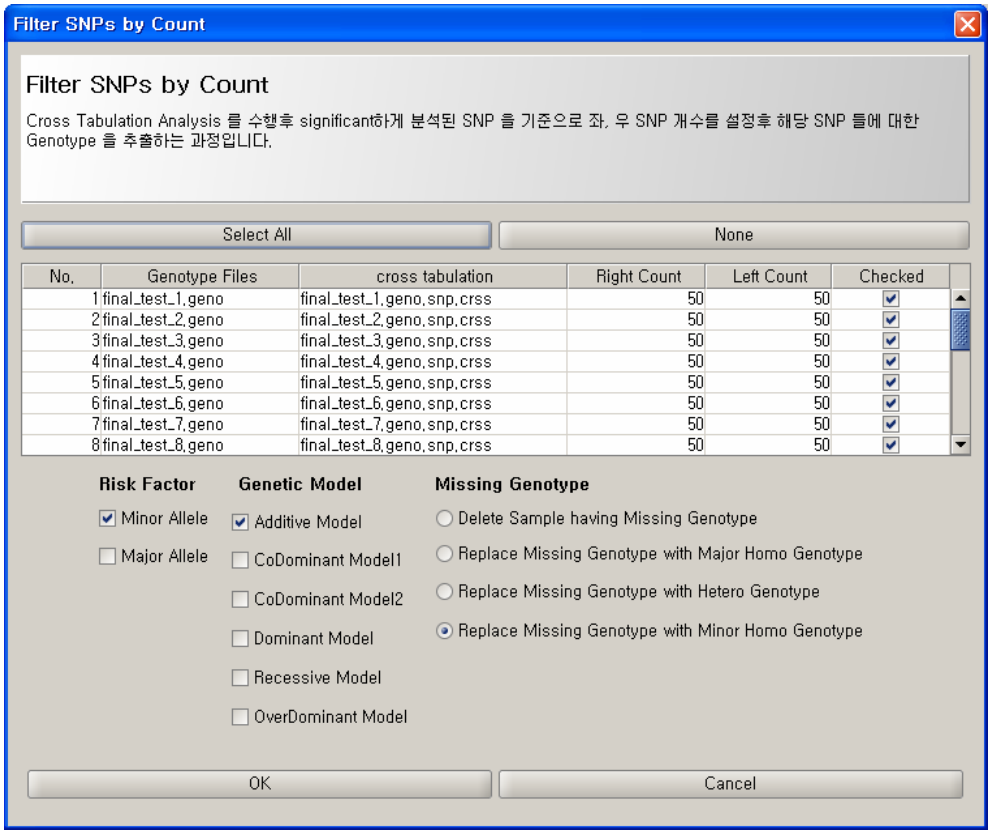
After completing the filtering task, PreProcess and annotation information extraction are automatically performed on the filtered genotype data and the result is added in project tree. (The result is added as *.SNP_ID.dis.filter.geno format in project tree.)

6.1.2. Filter SNPs by Count

It is possible to filter SNPs that are adjacent to the statistically significant SNPs by specifying number of adjacent SNPs. Click [Filter] > [Filter SNPs by Count] to show the setting window as in <Figure 6-2>. The result files from cross tabulation analysis (file extension is *.crss) and genotype files (extension is *.geno) are listed. Enter the proper numbers in the “Right Count” and “Left Count”. Click [OK] after setting "Risk Factor, "Genetic Model", and “Missing Genotype" to start the filtering process.

- Risk Factor: Risk Factor set in Cross Tabulation Analysis
  - Minor Allele / Major Allele
- Genetic Model: Analyzed model set in Cross Tabulation Analysis
  - Additive / Codominant1 / Codominant2 / Dominant / Recessive / Overdominant
- Missing Genotype: Missing genotype processing method

- Reserve Missing Genotype
- Replace Missing Genotype with Major Homo Genotype
- Replace Missing Genotype with Hetero Genotype
- Replace Missing Genotype with Minor Homo Genotype

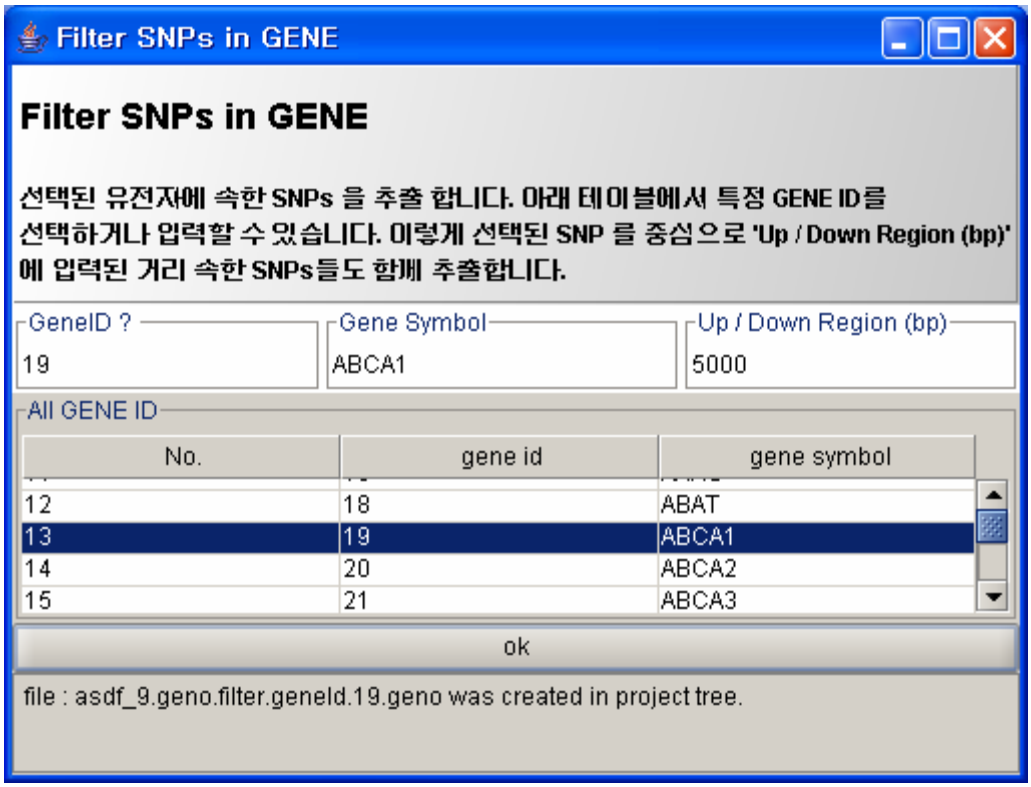


<Figure 6-2> SNP filtering by specifying number of adjacent SNPs

After completing the filtering task, PreProcess and annotation information extraction are automatically performed on the filtered genotype data and the result is added in project tree. (The result is added as *.SNP_ID.count.filter.geno format in project tree.)

6.2. Filter SNPs in GENE

SNPs in a specific genome area can be filtered. You can directly enter NCBI gene ID or gene symbol in the interface or select gene from the gene list. The table is sorted by gene ID. If you want to filter adjancent SNPs out of the boundary of the specificed gene together, just enter the base pair size in "Up / Down Region (bp)". Click [OK] for gene searching. If the search process is successful, the relevant *.geno file is automatically created and preprocessed.

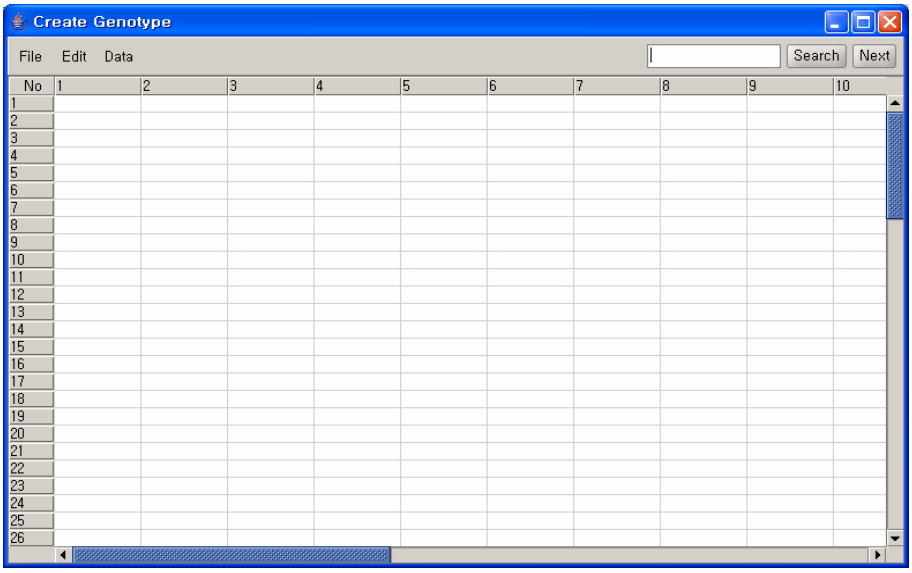


<Figure 6-3> Filter SNPs in GENE

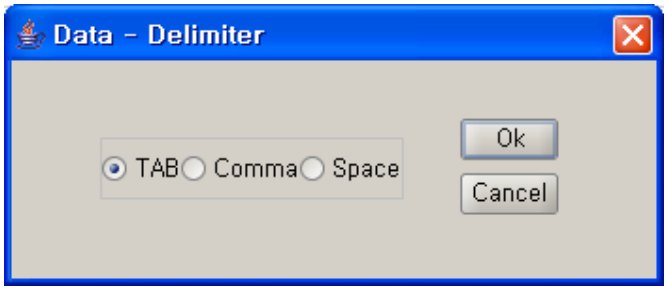
6.3. Data Edit

Users can create or modify data by using data editor. Click [Data] > [Data Edit] to show the empty editor as in <Figure 6-4>. Details are the following:

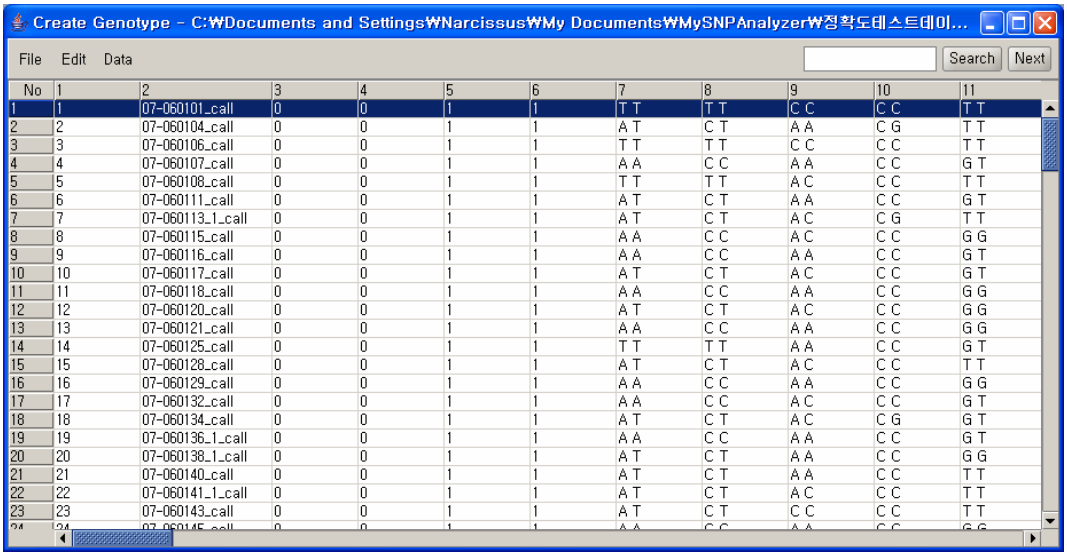
- Click [File] > [Open] and select file to edit.
- Click [OK] after selecting a text file delimiter as in <Figure 6-5>. <Figure 6-6> shows the contents of input data.
- Click [Edit] > [Space Insert] after selecting a row in the editor to add an empty row right above it. You can click [Edit] > [Remove] to remove the row.
- The [Edit] > [Copy] and [Edit] > [Paste] functions are the same as the Copy&Paste function of a typical data editor and [CTRL+C] and [CTRL+V] can be used as well.
- Click [Edit] > [Insert] to insert a copied contents as a new line.
- Click [Edit] > [Cut] to copy and remove the selected item from the editor. [CTRL+X] does the same function.



<Figure 6-4> Empty data editor

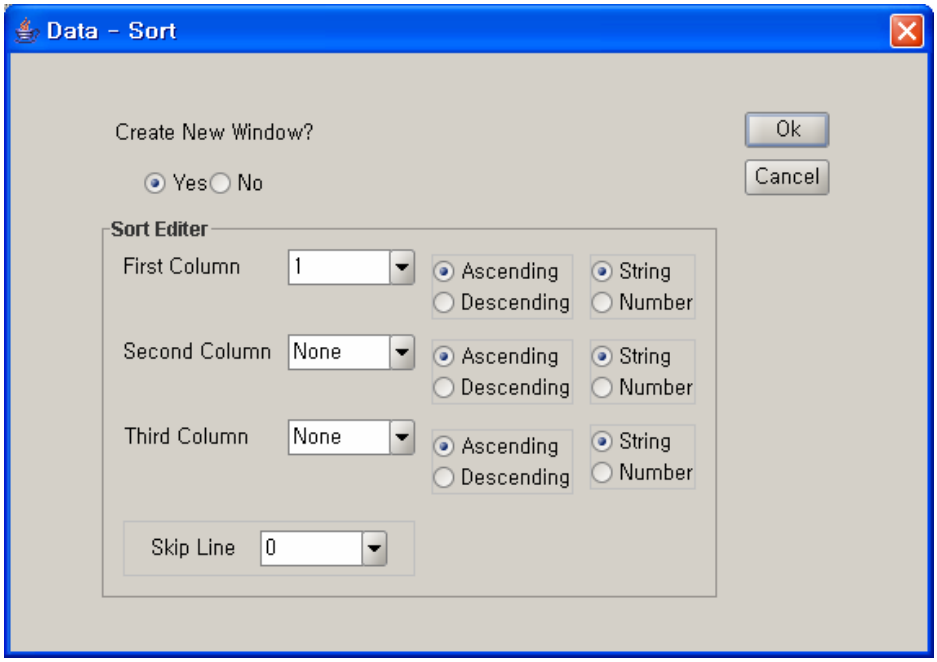


<Figure 6-5> Text file delimiter



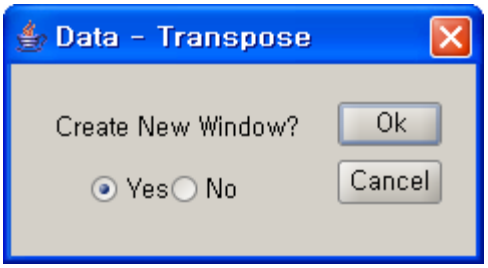
<Figure 6-6> Input data

- Click [Edit] > [Delete] to erase the selected contents without copying.
- Click [Data] > [Sort] to show the interface where you can sort the contents in the editor as in <Figure 6-7>. Select "Yes" to display the sorted result in a new window and select "No" to replace the existing contents with the sorted result. If you set "Skip Line" to "1", it sorts the data except for the first row of the data. Click [OK] after specifying the sorting options.



<Figure 6-7> Sorting options

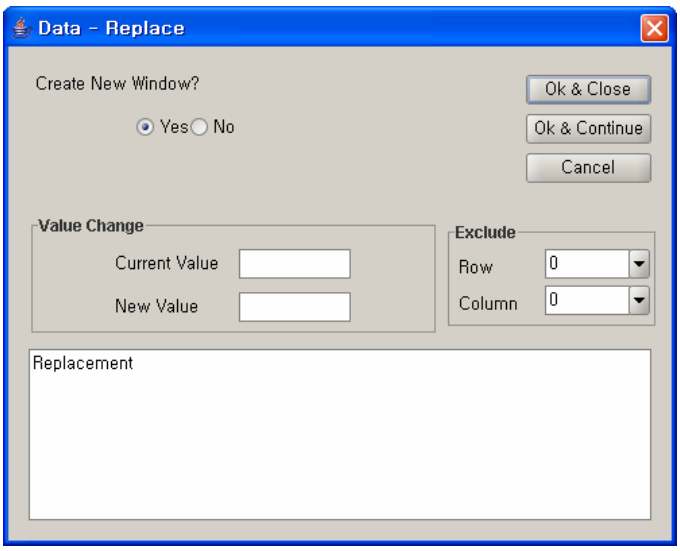
- Click [Data] > [Transpose] to transpose rows and columns of the data. Select “Yes” in <Figure 6-8> to create a new window and click [OK].



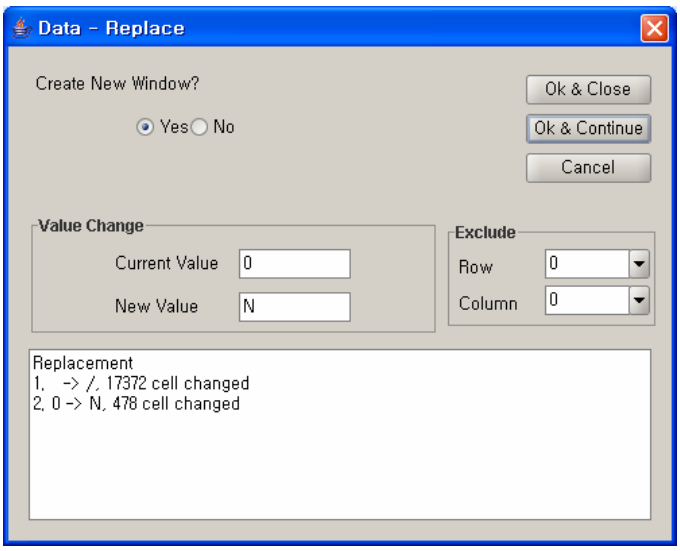
<Figure 6-8> Create a new window

- Click [Data] > [Replace] to replace a specific content of the data with other values. <Figure 6-9> is a replacement interface.
- Select “Yes” to create a new window and select “No” to replace the existing data.

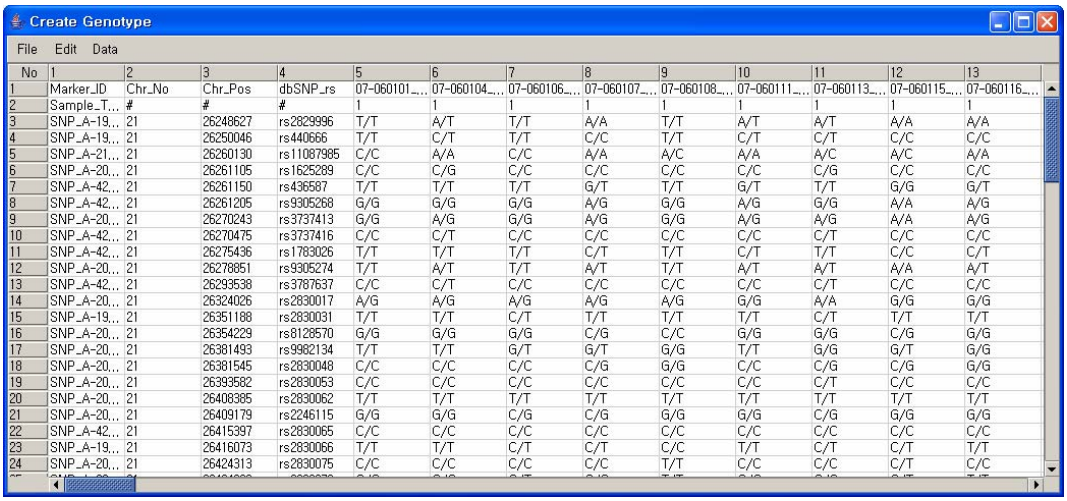
- If you specify "Row" to 2 and "Column" to 3 in "Exclude", the replacement will be executed except for 1 to 2 rows and 1 to 3 columns of the original data.
- Click [OK & Close] to execute the replacement and close the interface. Click [OK & Continue] to execute the replacement with the replacement history as in <Figure 6-10>. You can consecutively specify other replacement task.
- <Figure 6-11> is the replacement result.



<Figure 6-9> Replacement interface



<Figure 6-10> Replacement history



No	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Marker_ID	Chr_No	Chr_Pos	dbSNP_rs	07-060101_...	07-060104_...	07-060106_...	07-060107_...	07-060108_...	07-060111_...	07-060113_...	07-060115_...	07-060116_...
2	Sample_T...	#	#	#	1	1	1	1	1	1	1	1	1
3	SNP_A-19...	21	26248627	rs2829996	T/T	A/T	T/T	A/A	T/T	A/T	A/T	A/A	A/A
4	SNP_A-19...	21	26250046	rs440666	T/T	C/T	T/T	C/C	T/T	C/T	C/T	C/C	C/C
5	SNP_A-21...	21	26260130	rs11087985	C/C	A/A	C/C	A/A	A/C	A/A	A/C	A/C	A/A
6	SNP_A-20...	21	26261105	rs1625289	C/C	C/G	C/C	C/C	C/C	C/C	C/G	C/C	C/C
7	SNP_A-42...	21	26261150	rs436587	T/T	T/T	T/T	G/T	T/T	G/T	T/T	G/G	G/T
8	SNP_A-42...	21	26261205	rs9305268	G/G	G/G	G/G	A/G	G/G	A/G	G/G	A/A	A/G
9	SNP_A-20...	21	26270243	rs3737413	G/G	A/G	G/G	A/G	G/G	A/G	A/G	A/A	A/G
10	SNP_A-42...	21	26270475	rs3737416	C/C	C/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
11	SNP_A-42...	21	26275436	rs1783026	T/T	T/T	T/T	C/T	T/T	C/T	T/T	C/C	C/T
12	SNP_A-20...	21	26278851	rs9305274	T/T	A/T	T/T	A/T	T/T	A/T	A/T	A/A	A/T
13	SNP_A-42...	21	26293538	rs3787637	C/C	C/T	C/C	C/C	C/C	C/C	C/T	C/C	C/C
14	SNP_A-20...	21	26324026	rs2830017	A/G	A/G	A/G	A/G	A/G	G/G	A/A	G/G	G/G
15	SNP_A-19...	21	26351188	rs2830031	T/T	T/T	C/T	T/T	T/T	T/T	C/T	T/T	T/T
16	SNP_A-20...	21	26354229	rs8128570	G/G	G/G	G/G	C/G	C/C	G/G	G/G	C/G	G/G
17	SNP_A-20...	21	26381493	rs9982134	T/T	T/T	G/T	G/T	G/G	T/T	G/G	G/T	G/G
18	SNP_A-20...	21	26381545	rs2830048	C/C	C/C	C/C	C/G	G/G	C/C	C/G	C/G	G/G
19	SNP_A-20...	21	26393582	rs2830053	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C/C	C/C
20	SNP_A-20...	21	26408385	rs2830062	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T
21	SNP_A-20...	21	26409179	rs2246115	G/G	G/G	C/G	C/G	G/G	G/G	C/G	G/G	G/G
22	SNP_A-42...	21	26415397	rs2830065	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
23	SNP_A-19...	21	26416073	rs2830066	T/T	T/T	C/T	C/T	C/C	T/T	C/T	C/T	T/T
24	SNP_A-20...	21	26424313	rs2830075	C/C	C/C	C/C	C/C	T/T	C/C	C/C	C/T	C/C

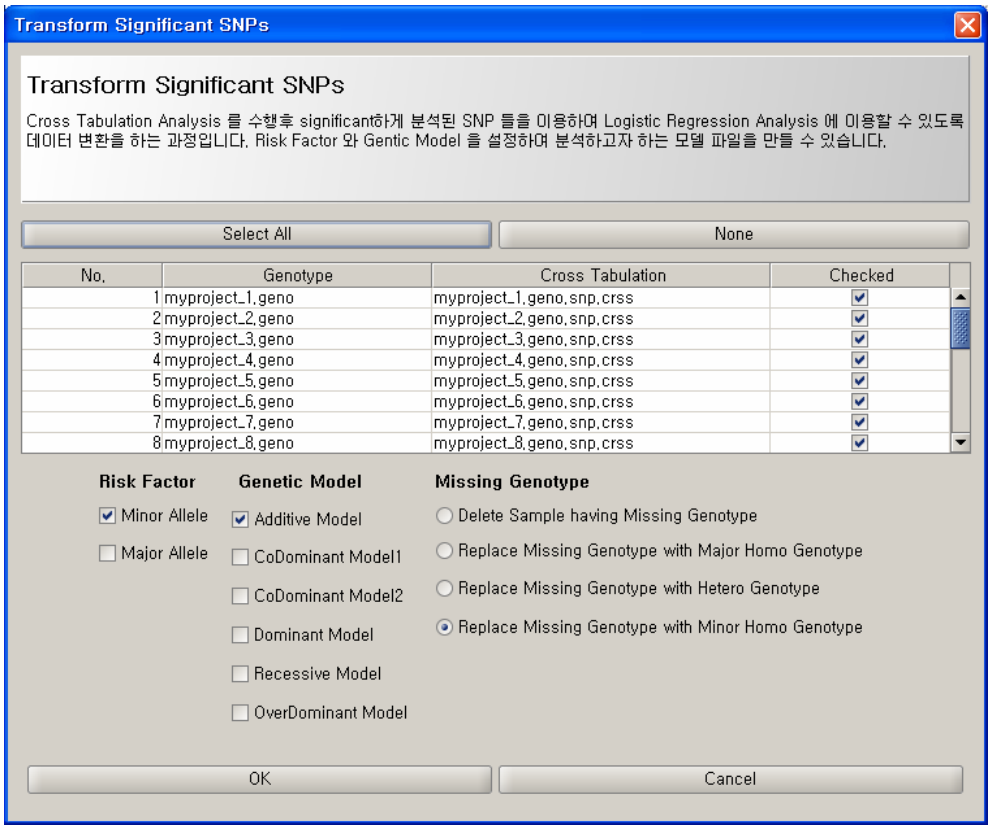
<Figure 6-11> Result of replacement

6.4. Transform

6.4.1. Transform Significant SNPs

Data transformation is required for losgistic regression analysis. Three different genotypes are transformed to “0”, “1” or “2” according to the type of risk allele and genetic model. If there is no transformed data (i.e model data) to analyze when performing the logistic regression, it automatically executes data trasnsformation process. Click [Transform] > [Transform Significant SNPs] to show a window as in <Figure 6-12>. Genotype files in project tree are listed in the “Genotype” list and the relevant files of cross tabulation analysis are listed in the “Cross Tabulation” list. Select files and set “Risk Factor”, “Genetic Model” and “Missing Genotype” for data transformation.

- Risk Factor: Risk Factor set in Cross Tabulation Analysis
  - Minor Allele / Major Allele
- Genetic Model: Analyzed model set in Cross Tabulation Analysis
  - Additive / Codominant1 / Codominant2 / Dominant / Recessive / Overdominant
- Missing Genotype: Missing genotype processing method
  - Reserve Missing Genotype
  - Replace Missing Genotype with Major Homo Genotype
  - Replace Missing Genotype with Hetero Genotype
  - Replace Missing Genotype with Minor Homo Genotype



<Figure 6-12> Data transformation control interface with significant SNPs

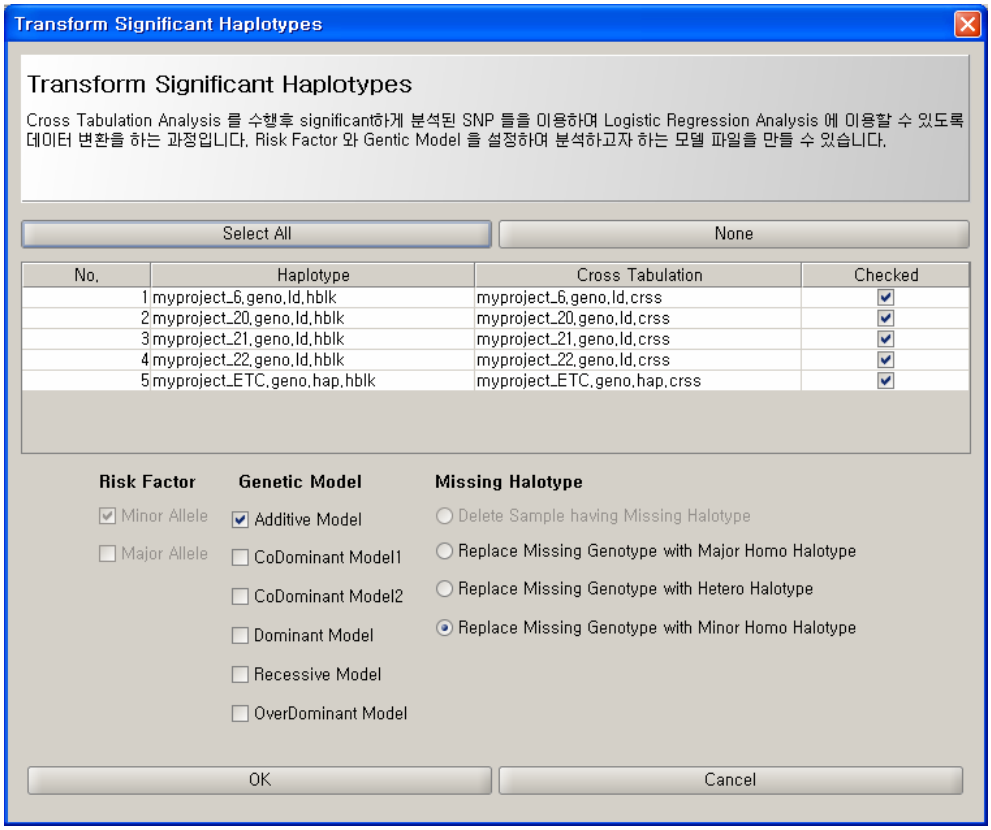
When data transformation is completed, the created model data is added in project tree.

6.4.2. Transform Significant Haplotypes

Users can implement logistic regression with haplotypes by transforming haplotype data into model data. If the significant haplotype is h1, then there are three different diplotypes: h1h1, h1h* and h*h*, where h* represents any of haplotypes other than h1. Three different genotypes are transformed to “0”, “1” or “2” according to the type of genetic model. If there is no transformed data (i.e model data) to analyze when performing the logistic regression, it automatically executes data transformation process. Click [Transfrom] > [Transform Signficant Haplotypes] to show a window as in <Figure 6-13>. Haplotype files in project tree are listed in the “Haplotype” list and the relevant files of cross tabulation anlaysis are listed in the “Cross Tabulation” list. Select files and set “Genetic Model” and “Missing Genotype” for data transformation.

- Risk Factor: Risk Factor set in Cross Tabulation Analysis
- Genetic Model: Analyzed model set in Cross Tabulation Analysis
  - Additive / Codominant1 / Codominant2 / Dominant / Recessive / Overdominant

- Missing Genotype: Missing genotype processing method
  - Reserve Missing Genotype
  - Replace Missing Genotype with Major Homo Genotype
  - Replace Missing Genotype with Hetero Genotype
  - Replace Missing Genotype with Minor Homo Genotype

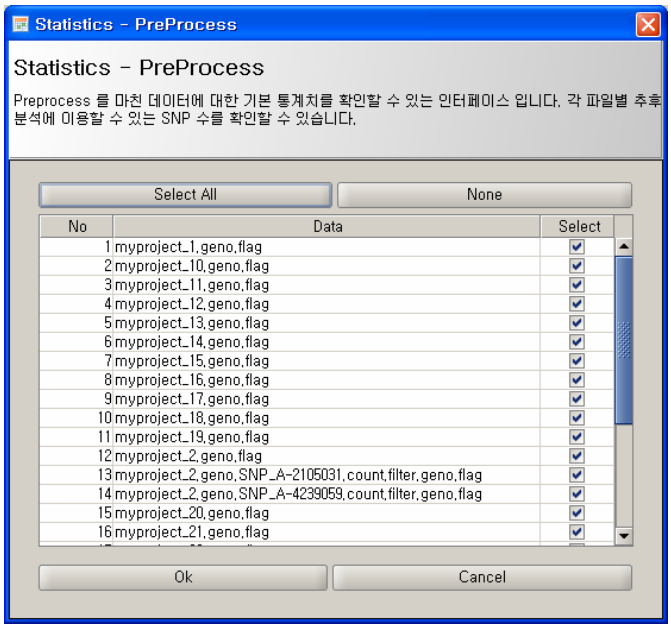


<Figure 6-13> transformation control interface with significant haplotypes

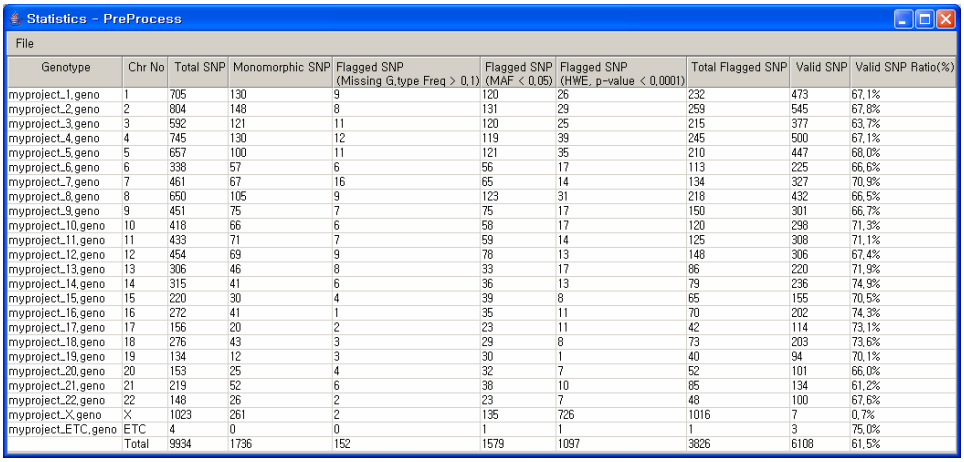
6.5. Statistics

6.5.1. PreProcess Statistics

Click [Statistics] > [PreProcess] to display a window where you can view the statistics for preprocessing result as in <Figure 6-14>. Click [OK] after selecting data to view, and the statistics result is created as shown in <Figure 6-15>. For the details about statistics contents, please refer to **Chapter 3 PreProcess.**



<Figure 6-14> Selection of preprocessing result

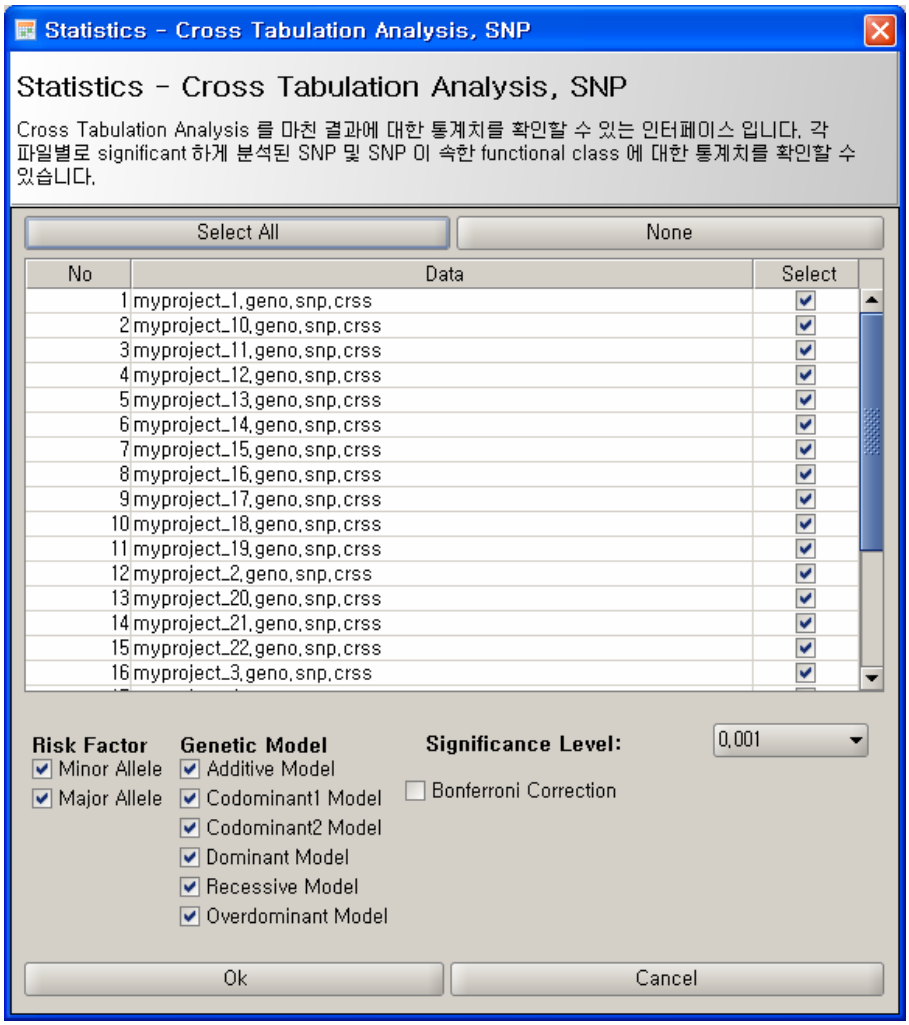


Genotype	Chr No	Total SNP	Monomorphic SNP	Flagged SNP (Missing G.type Freq > 0.1)	Flagged SNP (MAF < 0.05)	Flagged SNP (HWE, p-value < 0.0001)	Total Flagged SNP	Valid SNP	Valid SNP Ratio(%)
myproject_1.geno	1	705	130	9	120	26	232	473	67.1%
myproject_2.geno	2	804	148	8	131	29	259	545	67.8%
myproject_3.geno	3	592	121	11	120	25	215	377	63.7%
myproject_4.geno	4	745	130	12	119	39	245	500	67.1%
myproject_5.geno	5	657	100	11	121	35	210	447	68.0%
myproject_6.geno	6	338	57	6	56	17	113	225	66.6%
myproject_7.geno	7	461	67	16	65	14	134	327	70.9%
myproject_8.geno	8	650	105	9	123	31	218	432	66.5%
myproject_9.geno	9	451	75	7	75	17	150	301	66.7%
myproject_10.geno	10	418	66	6	58	17	120	298	71.3%
myproject_11.geno	11	433	71	7	59	14	125	308	71.1%
myproject_12.geno	12	454	69	9	78	13	148	306	67.4%
myproject_13.geno	13	306	46	8	33	17	86	220	71.9%
myproject_14.geno	14	315	41	6	36	13	79	236	74.9%
myproject_15.geno	15	220	30	4	39	8	65	155	70.5%
myproject_16.geno	16	272	41	1	35	11	70	202	74.3%
myproject_17.geno	17	156	20	2	23	11	42	114	73.1%
myproject_18.geno	18	276	43	3	29	8	73	203	73.6%
myproject_19.geno	19	134	12	3	30	1	40	94	70.1%
myproject_20.geno	20	153	25	4	32	7	52	101	66.0%
myproject_21.geno	21	219	52	6	38	10	85	134	61.2%
myproject_22.geno	22	148	26	2	23	7	48	100	67.6%
myproject_X.geno	X	1023	261	2	135	726	1016	7	0.7%
myproject_ETC.geno	ETC	4	0	0	1	1	1	3	75.0%
	Total	9934	1736	152	1579	1097	3826	6108	61.5%

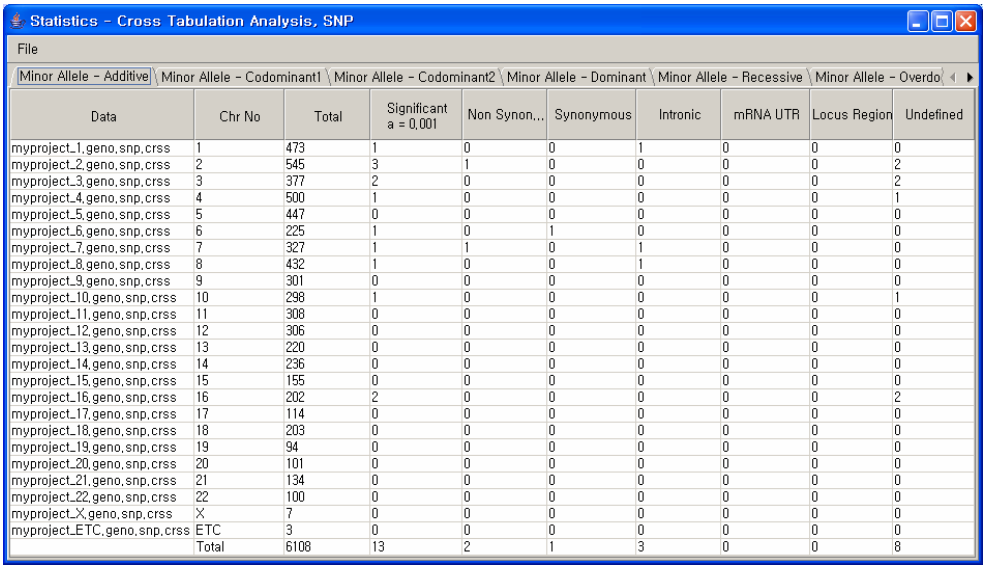
<Figure 6-15> Statistics result

6.5.2. Cross Tabulation Analysis Result Statistics

Click [Statistics] > [Cross Tabulation Analysis (SNP)] to show a window where you can view the statistics for cross tabulation analysis result as in <Figure 6-16>. Click [OK] after selecting data you want to view and set risk factor, genetic model, significance level and multiple test correction in <Figure 6-16>. The details about the statistical contents of each table, please refer to Chapter 4 4.1 Cross Tabulation Analysis using SNP.



<Figure 6-16> Selection of cross tabulation analysis result with SNPs



Data	Chr No	Total	Significant a = 0,001	Non Synon...	Synonymous	Intronic	mRNA UTR	Locus Region	Undefined
myproject_1,geno,snp,crss	1	473	1	0	0	1	0	0	0
myproject_2,geno,snp,crss	2	545	3	1	0	0	0	0	2
myproject_3,geno,snp,crss	3	377	2	0	0	0	0	0	2
myproject_4,geno,snp,crss	4	500	1	0	0	0	0	0	1
myproject_5,geno,snp,crss	5	447	0	0	0	0	0	0	0
myproject_6,geno,snp,crss	6	225	1	0	1	0	0	0	0
myproject_7,geno,snp,crss	7	327	1	1	0	1	0	0	0
myproject_8,geno,snp,crss	8	432	1	0	0	1	0	0	0
myproject_9,geno,snp,crss	9	301	0	0	0	0	0	0	0
myproject_10,geno,snp,crss	10	298	1	0	0	0	0	0	1
myproject_11,geno,snp,crss	11	308	0	0	0	0	0	0	0
myproject_12,geno,snp,crss	12	306	0	0	0	0	0	0	0
myproject_13,geno,snp,crss	13	220	0	0	0	0	0	0	0
myproject_14,geno,snp,crss	14	236	0	0	0	0	0	0	0
myproject_15,geno,snp,crss	15	155	0	0	0	0	0	0	0
myproject_16,geno,snp,crss	16	202	2	0	0	0	0	0	2
myproject_17,geno,snp,crss	17	114	0	0	0	0	0	0	0
myproject_18,geno,snp,crss	18	203	0	0	0	0	0	0	0
myproject_19,geno,snp,crss	19	94	0	0	0	0	0	0	0
myproject_20,geno,snp,crss	20	101	0	0	0	0	0	0	0
myproject_21,geno,snp,crss	21	134	0	0	0	0	0	0	0
myproject_22,geno,snp,crss	22	100	0	0	0	0	0	0	0
myproject_X,geno,snp,crss	X	7	0	0	0	0	0	0	0
myproject_ETC,geno,snp,crss	ETC	3	0	0	0	0	0	0	0
	Total	6108	13	2	1	3	0	0	8

<Figure 6-17> Statistics result

Chapter 7

Data Format

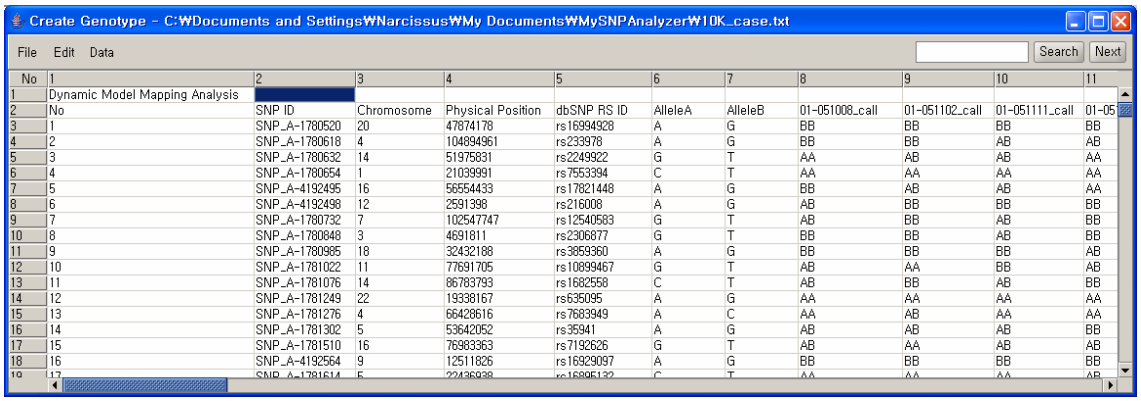
7. Data Format

7.1. Input Data Format

7.1.1. Affymetrix GeneChip GTYPE

Affymetrix GeneChip data should be in GTYPE format for using in SNPAnalyzer-Pro. GTYPE is a freely abailable software provided by Affymetrix Inc. You can download and install GCOS and GTYPE software free from Affymetrix homepage (http://www.affymetrix.com/products/software/index.affx). <Figure 7-1> shows an example of genotype format created using GTYPE, which can be recognized in SNPAnalyzer-Pro.

- First row
  - Algorithm name used by GTYPE software to extract genotype; Ex) Dynamic Model Mapping Analysis
- Second row
  - 1st column → **No**: Serial number of the SNP → reserved word
  - 2nd column → **SNP ID**: Probe set ID → reserved word
  - 3rd column → **Chromosome**: Chromosome number→ reserved word
  - 4th column → **Physical Position**: SNP position in chromosome → reserved word
  - 5th column → **dbSNP RS ID**: dbSNP #rs of the SNP → reserved word
  - 6th column → **AlleleA**: one allele of SNP → reserved word
  - 7th column → **AlleleB**: the other allele of SNP → reserved word
  - 8th column → 01-051008_**call**: Individual ID → only the “call” is a reserved word
  - Other columns → same as 8th column
- Third and other rows
  - The values corresponding to each column of the second row.
  - Individual genotype should be represented as “AA”, “AB”, “BB” or “NoCall”.



No	1	2	3	4	5	6	7	8	9	10	11
1	Dynamic Model Mapping Analysis										
2	No	SNP ID	Chromosome	Physical Position	dbSNP RS ID	AlleleA	AlleleB	01-051008_call	01-051102_call	01-051111_call	01-05
3	1	SNP_A-1780520	20	47874178	rs16994928	A	G	BB	BB	BB	BB
4	2	SNP_A-1780618	4	104894961	rs233978	A	G	BB	BB	AB	AB
5	3	SNP_A-1780632	14	51975831	rs2249922	G	T	AA	AB	AB	AA
6	4	SNP_A-1780654	1	21039991	rs7553394	C	T	AA	AA	AA	AA
7	5	SNP_A-4192495	16	56554433	rs17821448	A	G	BB	AB	AB	AA
8	6	SNP_A-4192498	12	2591398	rs216008	A	G	AB	BB	BB	BB
9	7	SNP_A-1780732	7	102547747	rs12540583	G	T	AB	BB	AB	BB
10	8	SNP_A-1780848	3	4691811	rs2306877	G	T	BB	BB	AB	BB
11	9	SNP_A-1780985	18	32432188	rs3859360	A	G	BB	BB	BB	AB
12	10	SNP_A-1781022	11	77691705	rs10899467	G	T	AB	AA	BB	AB
13	11	SNP_A-1781076	14	86783793	rs1682558	C	T	AB	BB	AB	BB
14	12	SNP_A-1781249	22	19338167	rs635095	A	G	AA	AA	AA	AA
15	13	SNP_A-1781276	4	66428616	rs7683949	A	C	AA	AB	AA	AA
16	14	SNP_A-1781302	5	53642052	rs35941	A	G	AB	AB	AB	BB
17	15	SNP_A-1781510	16	76983363	rs7192626	G	T	AB	AA	AB	AB
18	16	SNP_A-4192564	9	12511826	rs16929097	A	G	BB	BB	BB	BB

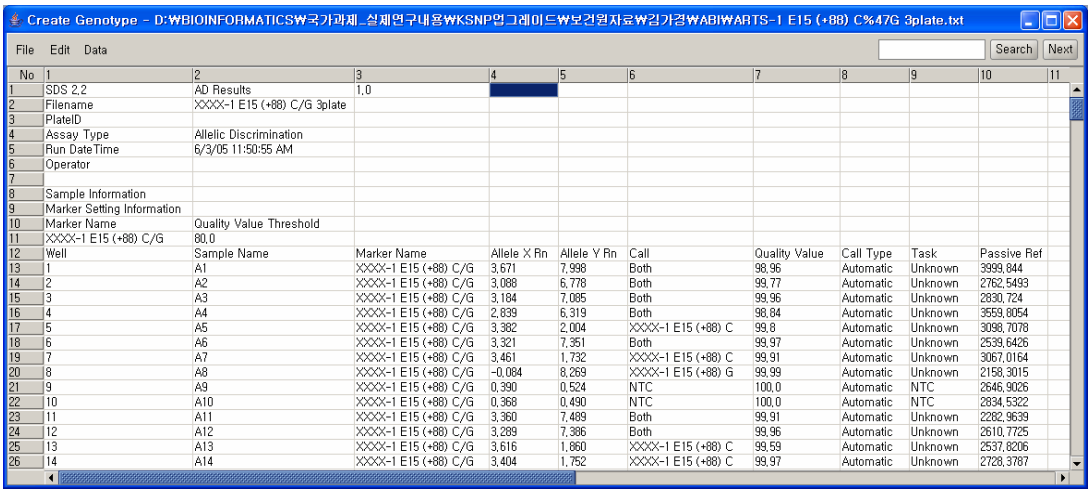
<Figure 7-1> Affymetrix GeneChip GTYPE data format

If the data format is different from the above, please refer to **Chapter 2 2.1.4 Genotype Import Data (Affymetrix GeneChip Data)**.

7.1.2. ABI TaqMan SNP Genotype

TagMan genotype data from ABI Inc. can be analyzed in SNPAnalyzer-Pro. <Figure 7-2> is an example of a genotype data.

- First column to 11th row
  - Headers describing the experiment
- 12th row
  - 1st column → **Well**: well number used in experiment → reserved word
  - 2nd column → **Sample Name**: sample identifier → reserved word
  - 3rd column → **Marker Name**: SNP identifier → reserved word
  - 4th column → Allele X Rn
  - 5th column → Allele Y Rn
  - 6th column → **Call**: genotype of sample → reserved word
  - 7th column → Quality Value
  - 8th column → Call Type
  - 9th column → Task
  - 10th column → Passive Ref
- 13th and other rows
  - The values corresponding to each column of the 12th row



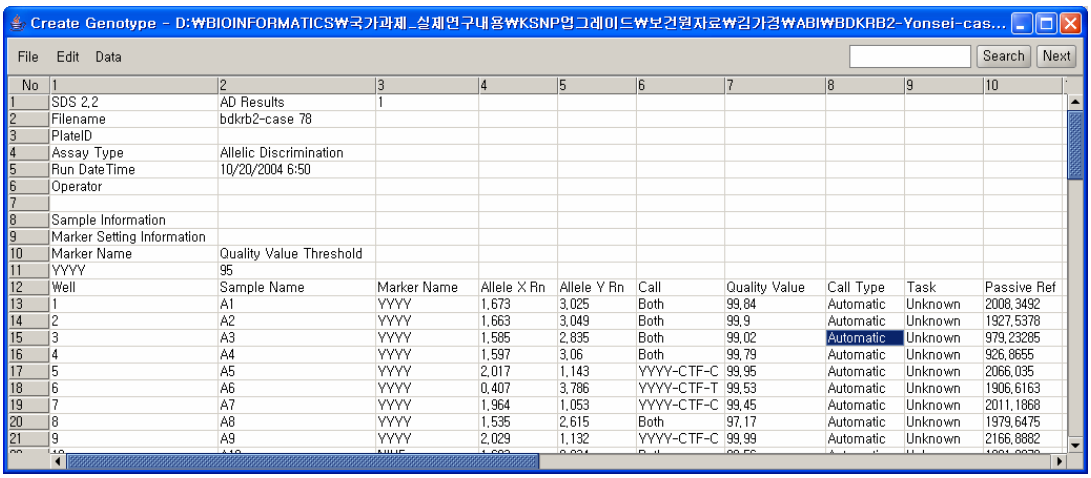
Create Genotype - D:₩BIOINFORMATICS₩국가과제_실제연구내용₩KSNP업그레이드₩보건원자료₩김가경₩ABI₩ARTS-1 E15 (+88) C%47G 3plate.txt

File Edit Data Search Next

No	1	2	3	4	5	6	7	8	9	10	11
1	SDS 2.2	AD Results	1.0								
2	Filename	XXXX-1 E15 (+88) C/G 3plate									
3	PlateID										
4	Assay Type	Allelic Discrimination									
5	Run DateTime	6/3/05 11:50:55 AM									
6	Operator										
7											
8	Sample Information										
9	Marker Setting Information										
10	Marker Name	Quality Value Threshold									
11	XXXX-1 E15 (+88) C/G	80.0									
12	Well	Sample Name	Marker Name	Allele X Rn	Allele Y Rn	Call	Quality Value	Call Type	Task	Passive Ref	
13	1	A1	XXXX-1 E15 (+88) C/G	3.671	7.998	Both	98.96	Automatic	Unknown	3999.844	
14	2	A2	XXXX-1 E15 (+88) C/G	3.088	6.778	Both	99.77	Automatic	Unknown	2762.5493	
15	3	A3	XXXX-1 E15 (+88) C/G	3.184	7.085	Both	99.96	Automatic	Unknown	2830.724	
16	4	A4	XXXX-1 E15 (+88) C/G	2.839	6.319	Both	98.84	Automatic	Unknown	3559.8054	
17	5	A5	XXXX-1 E15 (+88) C/G	3.382	2.004	XXXX-1 E15 (+88) C	99.8	Automatic	Unknown	3098.7078	
18	6	A6	XXXX-1 E15 (+88) C/G	3.321	7.351	Both	99.97	Automatic	Unknown	2539.6426	
19	7	A7	XXXX-1 E15 (+88) C/G	3.461	1.732	XXXX-1 E15 (+88) C	99.91	Automatic	Unknown	3067.0164	
20	8	A8	XXXX-1 E15 (+88) C/G	-0.084	8.269	XXXX-1 E15 (+88) G	99.99	Automatic	Unknown	2158.3015	
21	9	A9	XXXX-1 E15 (+88) C/G	0.390	0.524	NTC	100.0	Automatic	NTC	2646.9026	
22	10	A10	XXXX-1 E15 (+88) C/G	0.368	0.490	NTC	100.0	Automatic	NTC	2834.5322	
23	11	A11	XXXX-1 E15 (+88) C/G	3.360	7.489	Both	99.91	Automatic	Unknown	2282.9639	
24	12	A12	XXXX-1 E15 (+88) C/G	3.289	7.386	Both	99.96	Automatic	Unknown	2610.7725	
25	13	A13	XXXX-1 E15 (+88) C/G	3.616	1.860	XXXX-1 E15 (+88) C	99.59	Automatic	Unknown	2537.8206	
26	14	A14	XXXX-1 E15 (+88) C/G	3.404	1.752	XXXX-1 E15 (+88) C	99.97	Automatic	Unknown	2728.3787	

<Figure 7-2> ABI TaqMan SNP genotype format

<Figure 7-3> is another data format that SNPAnalyzer-Pro can automatically recognize.



Create Genotype - D:₩BIOINFORMATICS₩국가과제_실제연구내용₩KSNP업그레이드₩보건원자료₩김가경₩ABI₩BDKRB2-Yonsei-cas...

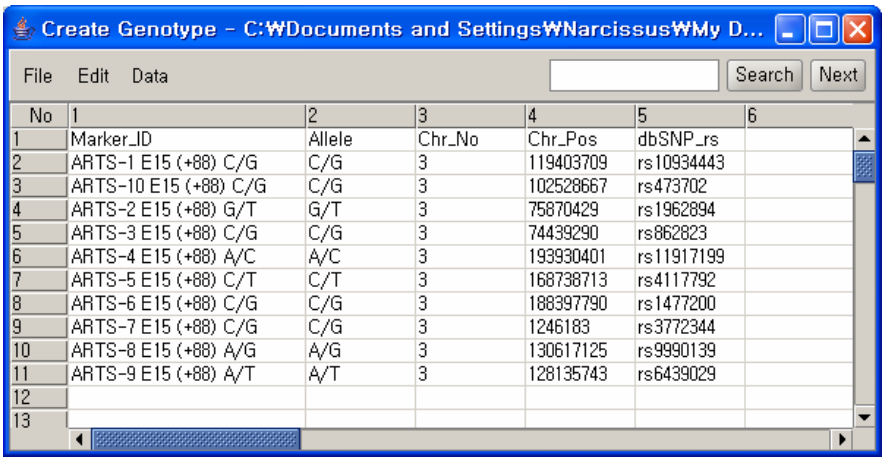
File Edit Data Search Next

No	1	2	3	4	5	6	7	8	9	10
1	SDS 2.2	AD Results	1							
2	Filename	bdkrb2-case 78								
3	PlateID									
4	Assay Type	Allelic Discrimination								
5	Run DateTime	10/20/2004 6:50								
6	Operator									
7										
8	Sample Information									
9	Marker Setting Information									
10	Marker Name	Quality Value Threshold								
11	YYYY	95								
12	Well	Sample Name	Marker Name	Allele X Rn	Allele Y Rn	Call	Quality Value	Call Type	Task	Passive Ref
13	1	A1	YYYY	1.673	3.025	Both	99.84	Automatic	Unknown	2008.3492
14	2	A2	YYYY	1.663	3.049	Both	99.9	Automatic	Unknown	1927.5378
15	3	A3	YYYY	1.585	2.835	Both	99.02	Automatic	Unknown	979.23285
16	4	A4	YYYY	1.597	3.06	Both	99.79	Automatic	Unknown	926.8655
17	5	A5	YYYY	2.017	1.143	YYYY-CTF-C	99.95	Automatic	Unknown	2066.035
18	6	A6	YYYY	0.407	3.786	YYYY-CTF-T	99.53	Automatic	Unknown	1906.6163
19	7	A7	YYYY	1.964	1.053	YYYY-CTF-C	99.45	Automatic	Unknown	2011.1868
20	8	A8	YYYY	1.535	2.615	Both	97.17	Automatic	Unknown	1979.6475
21	9	A9	YYYY	2.029	1.132	YYYY-CTF-C	99.99	Automatic	Unknown	2166.8882

<Figure 7-3> ABI TaqMan SNP Genotype format

The format of the markers' annotation file are as follows.

- Marker_ID: SNP ID
- Allele: two alleles separted by "/"
- Chr_No: chromosome in which SNP is located
- Chr_Pos: position of SNP in chromosome
- dbSNP_rs: dbSNP #rs of SNP

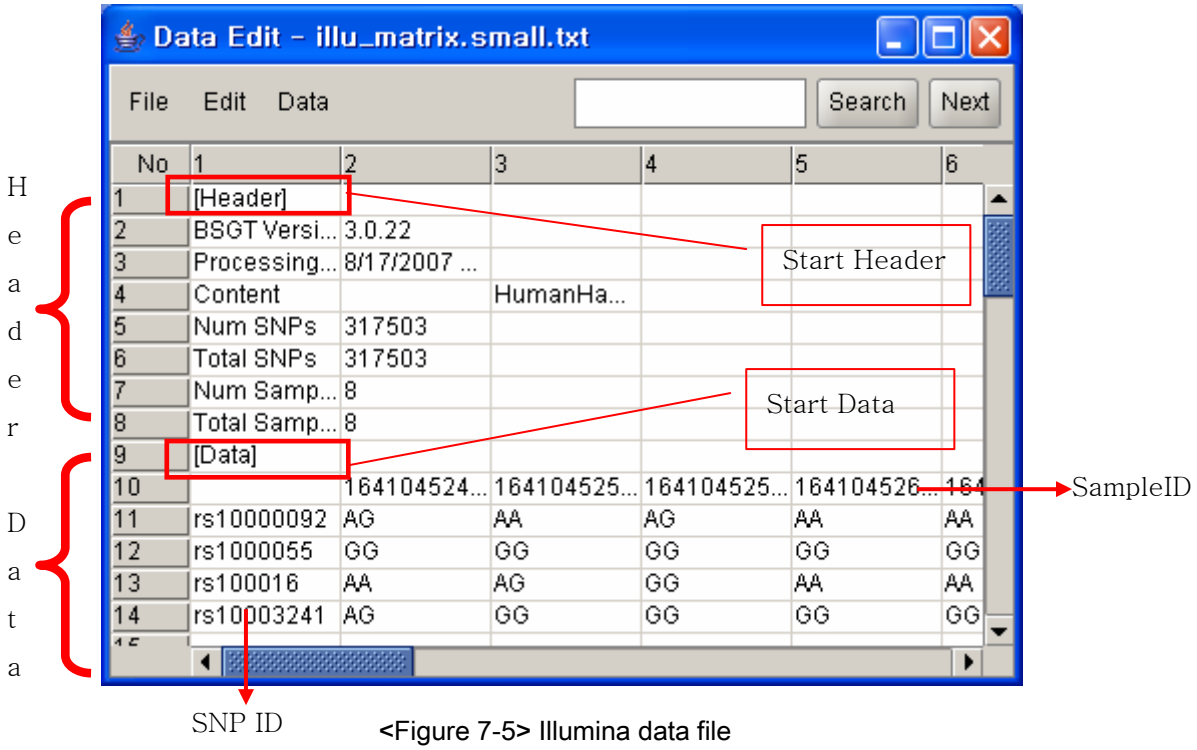


No	1	2	3	4	5	6
1	Marker_ID	Allele	Chr_No	Chr_Pos	dbSNP_rs	
2	ARTS-1 E15 (+88) C/G	C/G	3	119403709	rs10934443	
3	ARTS-10 E15 (+88) C/G	C/G	3	102528667	rs473702	
4	ARTS-2 E15 (+88) G/T	G/T	3	75870429	rs1962894	
5	ARTS-3 E15 (+88) C/G	C/G	3	74439290	rs862823	
6	ARTS-4 E15 (+88) A/C	A/C	3	193930401	rs11917199	
7	ARTS-5 E15 (+88) C/T	C/T	3	168738713	rs4117792	
8	ARTS-6 E15 (+88) C/G	C/G	3	188397790	rs1477200	
9	ARTS-7 E15 (+88) C/G	C/G	3	1246183	rs3772344	
10	ARTS-8 E15 (+88) A/G	A/G	3	130617125	rs9990139	
11	ARTS-9 E15 (+88) A/T	A/T	3	128135743	rs6439029	
12						
13						

<Figure 7-4> SNP marker annotation

7.1.3. Illumina SNP Genotype

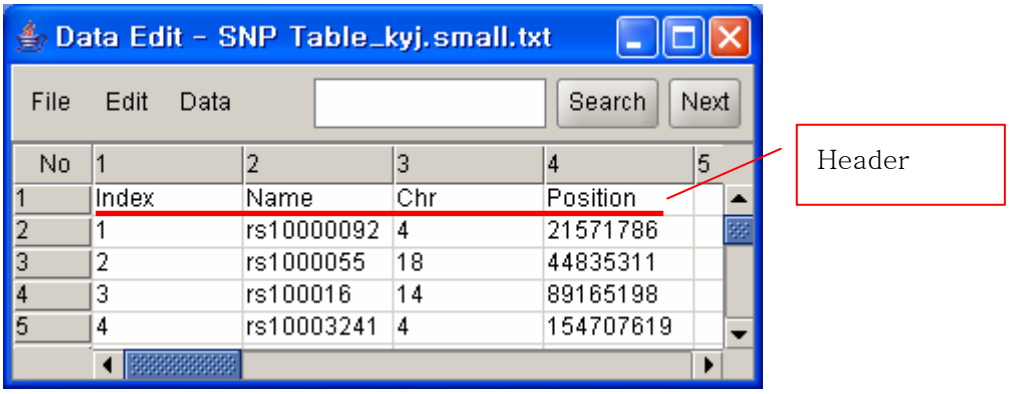
Two types of files are necessary for the analysis. One is genotype data and the other is SNP information data. Genotype data file is shown in <Figure 7-5>.



No	1	2	3	4	5	6
1	[Header]					
2	BSGT Versi...	3.0.22				
3	Processing...	8/17/2007 ...				
4	Content		HumanHa...			
5	Num SNPs	317503				
6	Total SNPs	317503				
7	Num Samp...	8				
8	Total Samp...	8				
9	[Data]					
10		164104524...	164104525...	164104525...	164104526...	164
11	rs10000092	AG	AA	AG	AA	AA
12	rs1000055	GG	GG	GG	GG	GG
13	rs100016	AA	AG	GG	AA	AA
14	rs10003241	AG	GG	GG	GG	GG

<Figure 7-5> Illumina data file

The SNP information file format is shown in <Figure 7-6>. “Name", “Chr", and “Position” are mandatory: “Name” for dbSNP ID, “Chr” for chromosome number, and “Position” for SNP position in chromosome.

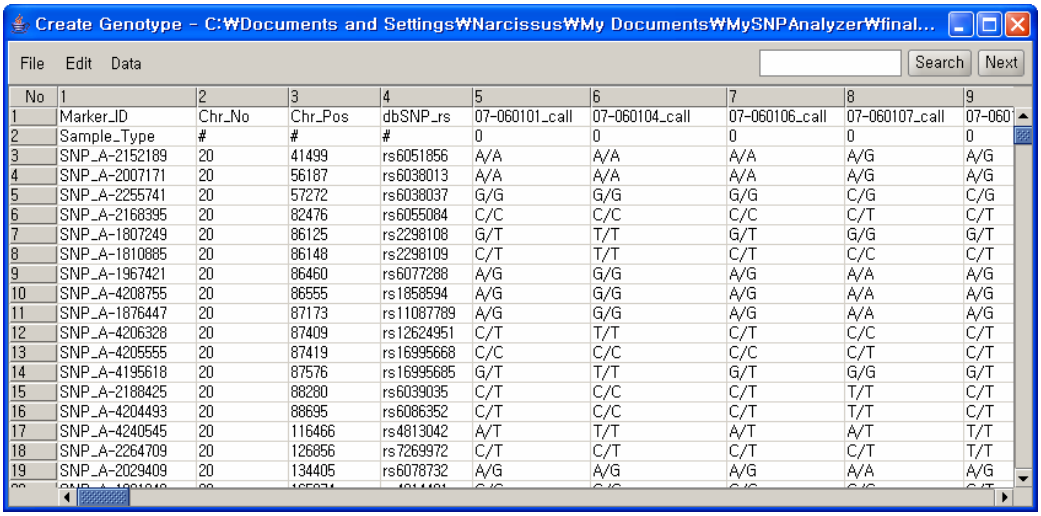


<Figure 7-6> Illumina SNP information file

7.1.4. SNPAnalyzer-Pro Specified Genotype (SNP To Sample) With SNP Annotation Format

The imported genotype data are automatically transformed into the format (extension *.geno) as in <Figure 7-7>. The below is description of each item.

- First row
  - 1st column → **Marker_ID** → reserved word
  - 2nd column → **Chr_No** → reserved word
  - 3rd column → **Chr_Pos**: SNP Position in Chromosome → reserved word
  - 4th column → **dbSNP_rs**: dbSNP #rs of SNP → reserved word
  - 5th column → Individual ID: sample ID
  - Other columns → same as the 5th column
- Second row
  - 1st column → **Sample_Type** → reserved word
  - 2nd to 4th columns → "#" → reserved word
  - 5th columns and others → control sample represented as "0" and case sample as "1"
- Third and other rows
  - The values corresponding to each column of the first row.
  - Individual genotype is represented as "A/A", "A/G" or "G/G". Missing genotype is codeded as "N/N".



No	1	2	3	4	5	6	7	8	9
1	Marker_ID	Chr_No	Chr_Pos	dbSNP_rs	07-060101_call	07-060104_call	07-060106_call	07-060107_call	07-060
2	Sample_Type	#	#	#	0	0	0	0	0
3	SNP_A-2152189	20	41499	rs6051856	A/A	A/A	A/A	A/G	A/G
4	SNP_A-2007171	20	56187	rs6038013	A/A	A/A	A/A	A/G	A/G
5	SNP_A-2255741	20	57272	rs6038037	G/G	G/G	G/G	C/G	C/G
6	SNP_A-2168395	20	82476	rs6055084	C/C	C/C	C/C	C/T	C/T
7	SNP_A-1807249	20	86125	rs2298108	G/T	T/T	G/T	G/G	G/T
8	SNP_A-1810885	20	86148	rs2298109	C/T	T/T	C/T	C/C	C/T
9	SNP_A-1967421	20	86460	rs6077288	A/G	G/G	A/G	A/A	A/G
10	SNP_A-4208755	20	86555	rs1858594	A/G	G/G	A/G	A/A	A/G
11	SNP_A-1876447	20	87173	rs11087789	A/G	G/G	A/G	A/A	A/G
12	SNP_A-4206328	20	87409	rs12624951	C/T	T/T	C/T	C/C	C/T
13	SNP_A-4205555	20	87419	rs16995668	C/C	C/C	C/C	C/T	C/T
14	SNP_A-4195618	20	87576	rs16995685	G/T	T/T	G/T	G/G	G/T
15	SNP_A-2188425	20	88280	rs6039035	C/T	C/C	C/T	T/T	C/T
16	SNP_A-4204493	20	88695	rs6086352	C/T	C/C	C/T	T/T	C/T
17	SNP_A-4240545	20	116466	rs4813042	A/T	T/T	A/T	A/T	T/T
18	SNP_A-2264709	20	126856	rs7269972	C/T	C/T	C/T	C/T	T/T
19	SNP_A-2029409	20	134405	rs6078732	A/G	A/G	A/G	A/A	A/G

<Figure 7-7> SNPAnalyzer-Pro specified genotype format

7.1.5. SNPAnalyzer-Pro Specified Genotype (SNP To Sample) Without SNP Annotation Format

This format contains only Marker ID and individual genotype information. Association analysis cannot be implemented with this format.

7.1.6. SNPAnalyzer-Pro Specified Genotype (Sample To SNP Format) With SNP Annotation Format

This format is the “SNPAnalyzer-Pro Specified Genotype (SNP to Sample Format) With SNP Annotation Format” format with rows and columns transposed

7.1.7. SNPAnalyzer-Pro Specified Genotype (Sample To SNP Format) Without SNP Annotation Format

This format is the “SNPAnalyzer-Pro Specified Genotype (SNP to Sample Format) Without SNP Annotation Format” format with rows and columns transposed

7.2. Annotation File Format

7.2.1. SNP Annotation File

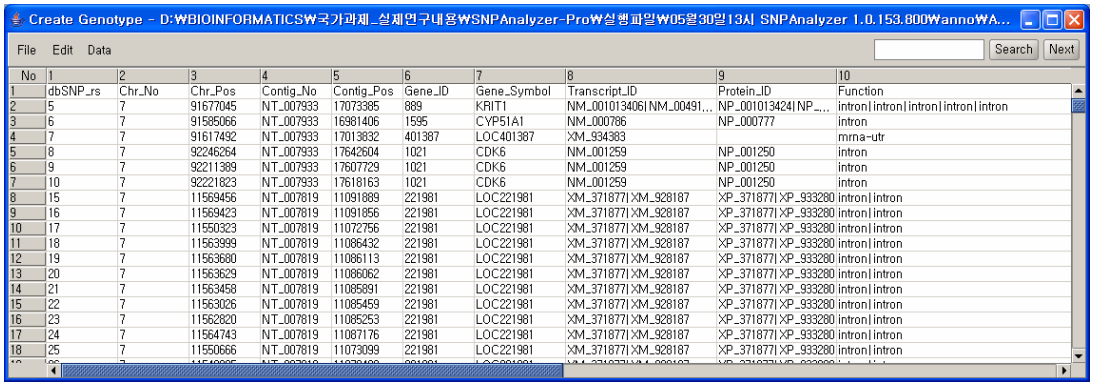
SNPAnlyzer-Pro provides the annotation information about SNPs and genes. <Figure 7-8> shows the annotation information about SNP. Descriptions for the SNP annotation information are the following:

- First row
  - 1st column→ dbSNP_rs: dbSNP #rs number

- 2nd column→ Chr_No: chromosome number to which SNP is mapped
- 3rd column → Chr_Pos: SNP position in chromosome
- 4th column→ Contig_No: Contig number to which SNP is mapped
- 5th column → Contig_Pos: SNP position in contig
- 6th column → Gene_ID: NCBI Gene ID to which SNP is mapped
- 7th column → Gene_Symbol: gene symbol
- 8th column → Transcript_ID: mRNA ID of the specified gene
- 9th column → Protein ID: protein ID of the specified mRNA
- 10th column → Function: functional class of the SNP

■ Second and other rows

- The values corresponding to each column of the first row.



No	1	2	3	4	5	6	7	8	9	10
1	dbSNP_rs	Chr_No	Chr_Pos	Contig_No	Contig_Pos	Gene_ID	Gene_Symbol	Transcript_ID	Protein_ID	Function
2	5	7	91677045	NT_007933	17073385	889	KRIT1	NM_001013406|NM_00491...	NP_001013424|NP_...	intron|intron|intron|intron|intron
3	6	7	91585066	NT_007933	16981406	1595	CYP51A1	NM_000786	NP_000777	intron
4	7	7	91617492	NT_007933	17013832	401387	LOC401387	XM_934383		mrna-utr
5	8	7	92246264	NT_007933	17642604	1021	CDK6	NM_001259	NP_001250	intron
6	9	7	92211389	NT_007933	17607729	1021	CDK6	NM_001259	NP_001250	intron
7	10	7	92221823	NT_007933	17618163	1021	CDK6	NM_001259	NP_001250	intron
8	15	7	11569456	NT_007819	11091889	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
9	16	7	11569423	NT_007819	11091856	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
10	17	7	11550323	NT_007819	11072756	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
11	18	7	11563999	NT_007819	11086432	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
12	19	7	11563680	NT_007819	11086113	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
13	20	7	11563629	NT_007819	11086062	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
14	21	7	11563458	NT_007819	11085891	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
15	22	7	11563026	NT_007819	11085459	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
16	23	7	11562820	NT_007819	11085253	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
17	24	7	11564743	NT_007819	11087176	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron
18	25	7	11550666	NT_007819	11073099	221981	LOC221981	XM_371877|XM_928187	XP_371877|XP_933280	intron|intron

<Figure 7-8> SNP snnotation information

7.2.2. Gene Annotation File

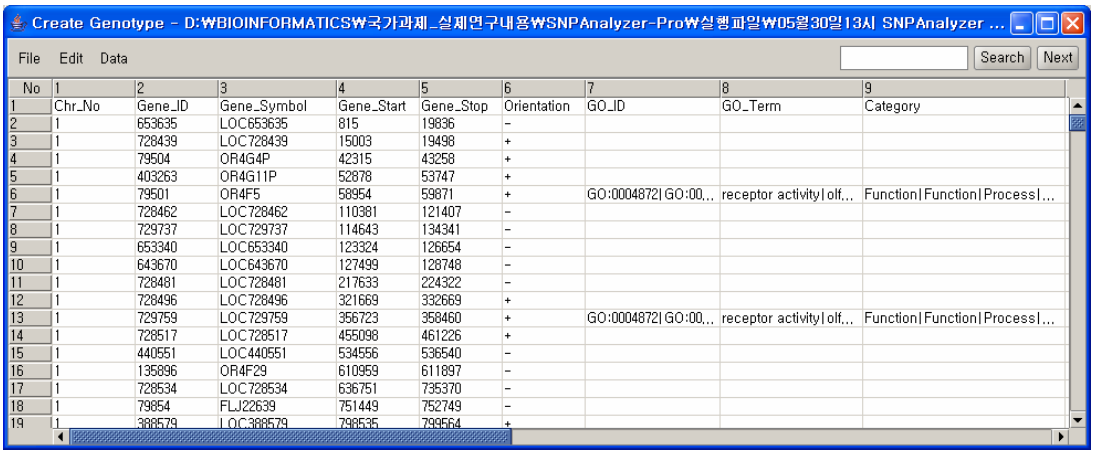
<Figure 7-9> shows the annotation information about genes. Descriptions for the gene annotation information are the following:

■ First row

- 1st column → Chr_No: Chromosome number to which gene is mapped
- 2nd column → Gene_ID: NCBI gene ID
- 3rd column → Gene_Symbol: gene symbol
- 4th column → Gene_Start: start position of gene in chromosome
- 5th column → Gene_Stop: stop position of gene in chromosome
- 6th column → Orientation: orientation of gene
- 7th column → GO_ID: gene ontology ID of gene
- 8th column → GO_Term: GO term
- 9th column → Category: GO category

■ Second and other columns

- The values corresponding to each column of the first row.



No	1	2	3	4	5	6	7	8	9
1	Chr_No	Gene_ID	Gene_Symbol	Gene_Start	Gene_Stop	Orientation	GO_ID	GO_Term	Category
2	1	653635	LOC653635	815	19836	-			
3	1	728439	LOC728439	15003	19498	+			
4	1	79504	OR4G4P	42315	43258	+			
5	1	403263	OR4G11P	52878	53747	+			
6	1	79501	OR4F5	58954	59871	+	GO:0004872| GO:00,..	receptor activity| olf,..	Function| Function| Process|...
7	1	728462	LOC728462	110381	121407	-			
8	1	729737	LOC729737	114643	134341	-			
9	1	653340	LOC653340	123324	126654	-			
10	1	643670	LOC643670	127499	128748	-			
11	1	728481	LOC728481	217633	224322	-			
12	1	728496	LOC728496	321669	332669	+			
13	1	729759	LOC729759	356723	358460	+	GO:0004872| GO:00,..	receptor activity| olf,..	Function| Function| Process|...
14	1	728517	LOC728517	455098	461226	+			
15	1	440551	LOC440551	534556	536540	-			
16	1	135896	OR4F29	610959	611897	-			
17	1	728534	LOC728534	636751	735370	-			
18	1	79854	FLJ22639	751449	752749	-			
19	1	388579	LOC388579	798535	799564	+			

<Figure 7-9> Gene annotation information

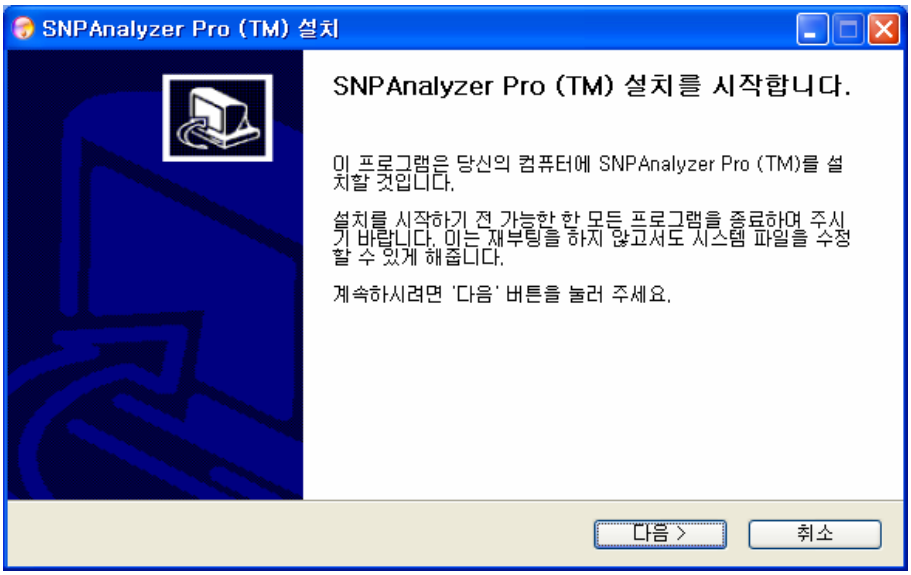
APPENDIX-A

Installation & Registration

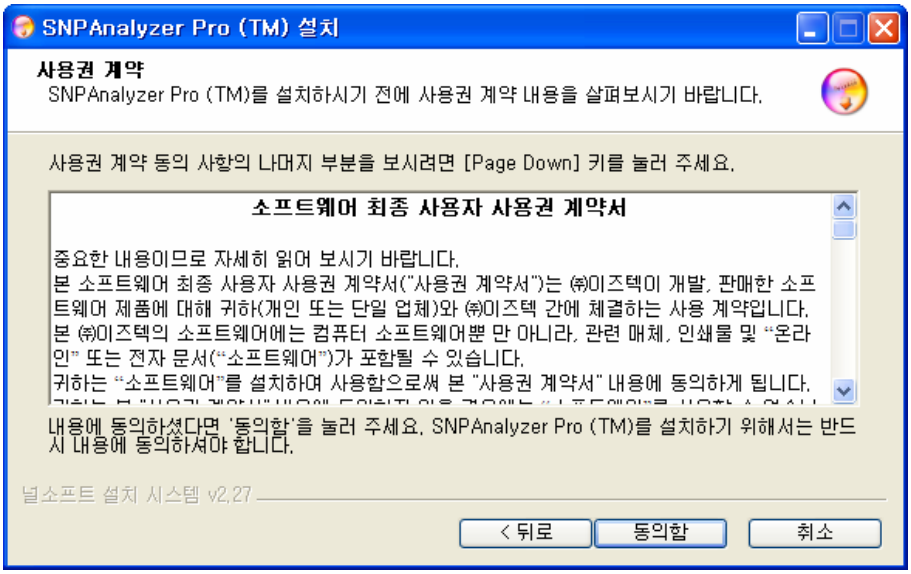
8. How to Install

Before you install SNPAnalyzer-Pro, check if your computer is connected to internet (SNPAnalyzer-Pro checks the license number).

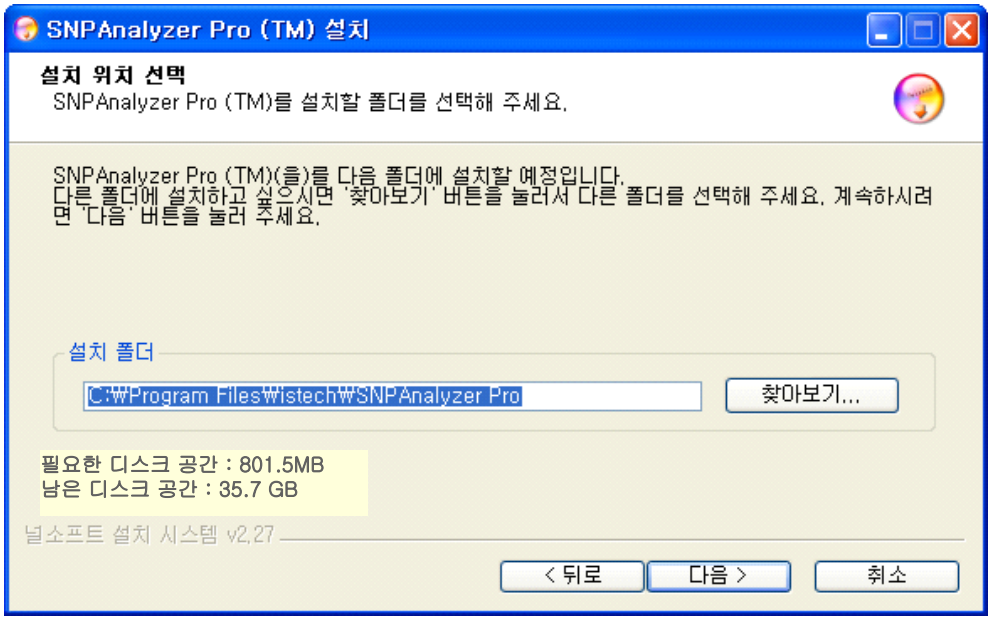
- Run the installation file (SNPAnalyzerPro-Setup.exe) you get from CD or downloading from homepage and a screen as below appears. Click [Next] to start installation.



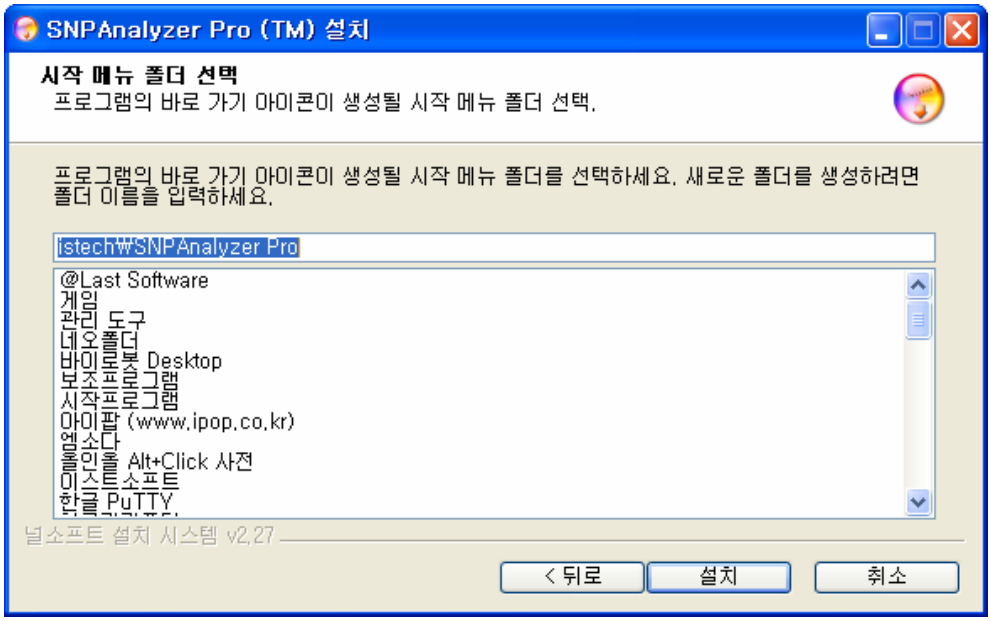
- The below is the license agreement. Click [Agree] to continue.



■ After specifying the location to install the software and click [Next].



■ Select the start menu and click [Install] to start the installation.



- When the software is installed successfully, click [Finish] to run SNPAnalyzer-Pro.

- In order to register the software online, click [Help]→[Online Register] in the main menu of the SNPAnalyzer-Pro.

APPENDIX-B

Algorithms

9. PreProcess

9.1. Hardy-Weinberg Equilibrium Test

It tests the if each SNP are in Hardy-Weinberg Equilibrium using chi-square test. The below are the basic table and calculation formula to perform the HWE test.

<Genotype Frequency Table>

	Genotype			
	Major Homo	**Hetero**	**Minor Homo**	**Sum**
Observed Freq	**O_1**	**O_2**	**O_3**	**n**
Exptected Freq	**E_1**	**E_2**	**E_3**	**n**

$$n = O_1 + O_2 + O_3$$

$$p = \frac{2O_1 + O_2}{2n}, q = \frac{O_2 + 2O_3}{2n}$$

$$E_1 = np^2,\ E_2 = 2npq,\ \ E_3 = nq^2$$

$$\mathrm{X}^2 = \sum_{i=1}^{3} \frac{\left(|O_i - E_i| - \frac{1}{2} \right)^2}{E_i},\ \ DOF = 1 \ \ \text{with Yates' Correction for Continuity}$$

9.2. Replace Missing Genotype

Missing genotypes can be replaced with one of the observed genotypes of each SNP. Replaceable genotypes are the following:

- Major homozygous genotype
- Minor homozygous genotype
- Heterozygous genotype

10. Cross Tabulation Analysis

10.1. Risk Factor / Genetic Model

You need to set risk factor and test model (genetic model) to implement case-control analysis. Major allele or minor allele can be specified as risk factor. Suppose the risk factor is "R" and wild factor is "W", two-by-two or three-by-two contingency table is used to perform case-control analysis. In the below tables, [RR], [RW], and [WW] show the number of genotypes observed in case sample. [R'R'], [R'W'], and [W'W'] show the number of genotypes observed in control sample.

- Additive Model:

	Case	Control
Risk	2*[RR] + [RW]	2*[R'R'] + [R'W']
Wild	[RW] + 2*[WW]	[R'W'] + 2*[W'W']

- Dominant Model:

	Case	Control
Risk	[RR] + [RW]	[R'R'] + [R'W']
Wild	[WW]	[W'W']

- Recessive Model:

	Case	Control
Risk	[RR]	[R'R']
Wild	[RW]+[WW]	[R'W'] + [W'W']

- Codominant Model:

	Case	Control
Genotype1	[RR]	[R'R']
Genotype2	[RW]	[R'W']
Genotype3	[WW]	[W'W']

■ Overdominant Model:

	Case	Control
Homo Genotype	[RR] + [WW]	[R'R'] + [W'W']
Hetero Genotype	[RW]	[R'W']

10.2. Odds Ratio, Attributable Risk (%), Population Attributable Risk (%)

Odds Ratios (OR) and its 95% confidence interval are calculated in the case-control analysis. Also, attributable risk percentage (AR%) and population attributable risk percentage (PAR%) are estimated in paralle with Odds Ratios. The below is the calculation for OR, AR% and PAR%.

<2x2 Contingency Table>

	Case	Control	Total
Risk Factor	**A**	**B**	**A+B**
Wild Factor	**C**	**D**	**C+D**
Total	**A+C**	**B+D**	**A+B+C+D**

$$OR = \frac{A/C}{B/D} = \frac{A \times D}{B \times C}: \text{ odds ratio} \rightarrow \text{likelihood of being sick}$$

$$100(1\text{-}\alpha)\%\ \text{CI of OR}: e^{\ln OR \pm (z_{\alpha/2} \times s_e)},\ s_e = \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}$$

Case – Control – Study Approximation

$$AR\% \approx \frac{OR-1}{OR} \times 100\%$$

$$PAR\% \approx \frac{p_{con} \times (OR-1)}{1 + p_{con} \times (OR-1)} \times 100\%, \qquad p_{con} = \frac{B}{D},$$

when $A, C << B, D$

10.3. Goodness of Fit Test & Likelihood Ratio Test

Tow types of chi-square test are used for the case-control analysis. One is goodness of fit test and the other is likelihood ratio test.

< 2x2 Contingency table for the observed value>

	Case	Control	Total
Risk Factor	O_{11}	O_{12}	O_{1+}
Wild Factor	O_{21}	O_{22}	O_{2+}
Total	O_{+1}	O_{+2}	m

< 2x2 Contingency table for the expected value>

	Case	Control	Total
Risk Factor	E_{11}	E_{12}	E_{1+}
Wild Factor	E_{21}	E_{22}	E_{2+}
Total	E_{+1}	E_{+2}	m

Goodness of Fit Test

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(|O_{ij}-E_{ij}|-\frac{1}{2}\right)^2}{E_{ij}},\ \mathrm{DOF}=(2-1)\times(2-1)\ ,\ with\ Yate's\ Correction$$

$$m = O_{11}+O_{12}+O_{21}+O_{22}$$

$$p_{i+}=\frac{O_{i+}}{m},\quad p_{+j}=\frac{O_{+j}}{m}$$

$$E_{11}=m\times p_{1+}\times p_{+1}$$

$$E_{12}=m\times p_{1+}\times p_{+2}$$

$$E_{21}=m\times p_{2+}\times p_{+1}$$

$$E_{22}=m\times p_{2+}\times p_{+2}$$

Likelihood Ratio Test

$$L_F=\sum_{i=1}^{2}\sum_{j=1}^{2}O_{ij}\times\ln\frac{O_{ij}}{m}$$

$$L_R=\sum_{i=1}^{2}\sum_{j=1}^{2}E_{ij}\times\ln\frac{E_{ij}}{m}$$

$$\chi^2=-2(L_R-L_F)=2\sum_{i=1}^{2}\sum_{j=1}^{2}O_{ij}\times\ln\frac{O_{ij}}{E_{ij}}\quad,\ \mathrm{DOF}=(2\text{-}1)\times(2\text{-}1)$$

11. Logistic Regression Analysis

11.1. Parameter Estimation

The analysis of the relationship between the response variable and explanatory variable is implemented using logistic regression when the response variable is binary type. The below is the formula describing the logistic model. For the logistic model, it estimates the approximated value of parameter (β) using the iteratively weighted least square method.

Logistic Regression Model

$$E\{Y_i\} = \pi_i = \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_i)}{1+\exp(\boldsymbol{\beta}'\mathbf{X}_i)}, \ where \quad i = 1, 2, \ldots n$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{bmatrix} \qquad \mathbf{X}_i = \begin{bmatrix} 1 \\ x_{1i} \\ \cdots \\ x_{p-1,i} \end{bmatrix}$$

$$Y_i = \{0,1\} \rightarrow Observed \ \ Sample \ \ Class$$

$$X_i = \{0,1,2\} \rightarrow Observed \ \ Genotype$$

There are four parameters to be set in the logistic regression analysis.

- Maximum Iteration: The number of times algorithm is performed to estimate parameter (β) when using the iteratively weighted least square method.
- Parameter Change Cutoff: The change in parameter (β) value when algorithm stops. Default value is 0.001.
- Classification Probability Cutoff: The probability that the observed class sample is determined to be class sample. You can select one of the following values: 0.1, 0.2, 0.3, 0.4, and 0.5. Default value is 0.5.
- Classification Power: The percentage of correct classification. Default value is 100%.

11.2. Classification Table

The below table shows the correctly or uncorrectly calssified sample count by logistic regression anlaysis.

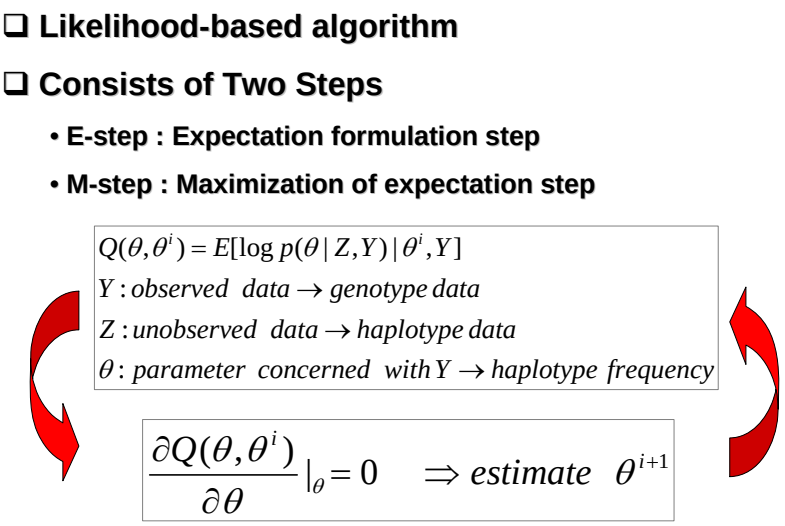
<2-Class Classification Table>

	Predicted Class			
Observed Class		**0**	**1**	**Classification Power**
	0	**A**	**B**	**A / (A+B)**
	1	**C**	**D**	**D / (C+D)**
	Overal	**A / (A+C)**	**D / (B+D)**	**(A+D) / (A+B+C+D)**

12. LD Analysis

12.1. Haplotype Estimation

Haplotypes can be reconstructed from genotypes using algorithms like EM or PL-EM. The EM algorithm estimates haplotypes ussing maximum likelihood estimation process. The PL-EM algorithm estimates haplotypes using the EM algorithm inside each block after dividing the entire SNPs into several blocks. It mearges adjacent blocks into one single block and re-construct merged haplotypes in that single block. The process goes on until the final one block remains. The Figure below shows the basic concept for EM algorithm.



12.2. Pairwise LD

There is a high possibility of strong linkage disequilibrium between SNPs located adjacent to each other. The degree of linkage disequilibrium relationship can be generally represented as indices like D' and r^2.

<Haplotype Frequency observed in adjacent SNP Pair>

Marker	SNP 2			Total
SNP 1		Allele 1	Allele 2	
	Allele 1	p_{11}	p_{12}	p_{1+}
	Allele 2	p_{21}	p_{22}	p_{2+}
Total		p_{+1}	p_{+2}	1

1) $D = p_{11} \times p_{22} - p_{12} p_{21}$

2) $$D' = \begin{cases} \dfrac{D}{\min(p_{1+} \times p_{+2}, p_{+1} \times p_{2+})} & if \quad D > 0 \\ \dfrac{D}{\min(p_{1+} \times p_{+1}, p_{+2} \times p_{2+})} & if \quad D < 0 \end{cases}$$

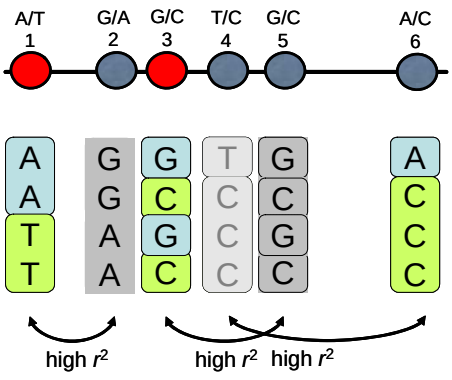
3) $$r = \frac{D}{(p_{1+} \times p_{2+} \times p_{+1} \times p_{+2})^{1/2}}$$

12.3. Tagging SNPs

Representative SNP that has strong correlation (r^2 > 0.8) with other SNPs is designated as pairwise tagging SNP.

❑ Genome-wide Tagging SNP

- **tagSNP : 다수의 SNP 에서 공통되게 나타나는 Allele 을 대표하는 SNP**
- **Pairwise SNP 간의 r^2 (상관관계지수) 이용**
- **Reference : Carlson et al., Am.J.Hum.Genet., 2004**



Tags

SNP 1
SNP 3
SNP 6

SNPs actually tested:

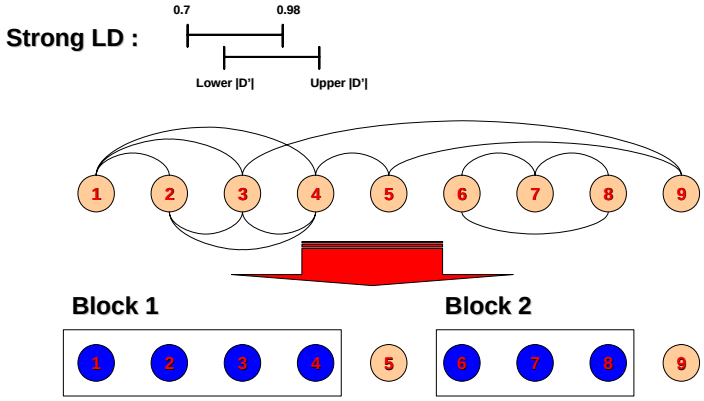
SNP 1
SNP 3
SNP 6
SNP 2
SNP 5
SNP 4

12.4. LD Block

Several SNPs that are in strong linkage disequilibrium can be bound into one LD block . For LD blocking, the Gabriel's method is used that is based on D'.

❑ **LD based 알고리듬**

- **LD Block : 서로 인접한 SNP 들중 강한 연관불평형 관계에 있는 SNP 집단**
- **|D'| 의 confidence interval(95%) 을 bootstrapping 이용해 계산**
- **Reference : Gabriel et al., Science, 2002**



12.5. Multi Allelic D'

The degree of linkage disequilibrium between contiguous LD blocks can be estimated with the multi allelic D'. The below are the table and calculation.

<Haplotype frequency observed in contiguous block pair>

Marker	Block 2						
Block 1		Allele 1	…	Allele j	…	Allele n	
	Allele 1	p_{11}	…	p_{1j}	…	p_{1n}	p_{1+}
	…	…	…	…	…	…	…
	Allele i	p_{i1}	…	p_{ij}	…	p_{in}	p_{i+}
	…	…	…	…	…	…	…
	Allele m	p_{m1}	…	p_{mj}	…	p_{mn}	p_{m+}
		p_{+1}	…	p_{+j}	…	p_{+n}	1

$$D' = \sum_{j=1}^{n}\sum_{i=1}^{m} p_{i+}\, p_{+j}\, |\, D'_{ij}\, |$$

$$D'_{ij} = \frac{D_{ij}}{D_{ij,\max}}$$

$$D_{ij} = p_{ij} - p_{i+}\, p_{+j}$$

$$D_{ij,\max} = \begin{cases} \min(p_{i+}\, p_{+j}, (1-p_{i+})(1-p_{+j}) & \text{if } D_{ij} < 0 \\ \min((1-p_{i+})\, p_{+j}, p_{i+}(1-p_{+j}) & \text{if } D_{ij} > 0 \end{cases}$$